Lernende Künstliche Intelligenz in der Rüstungskontrolle

Lück, Nico

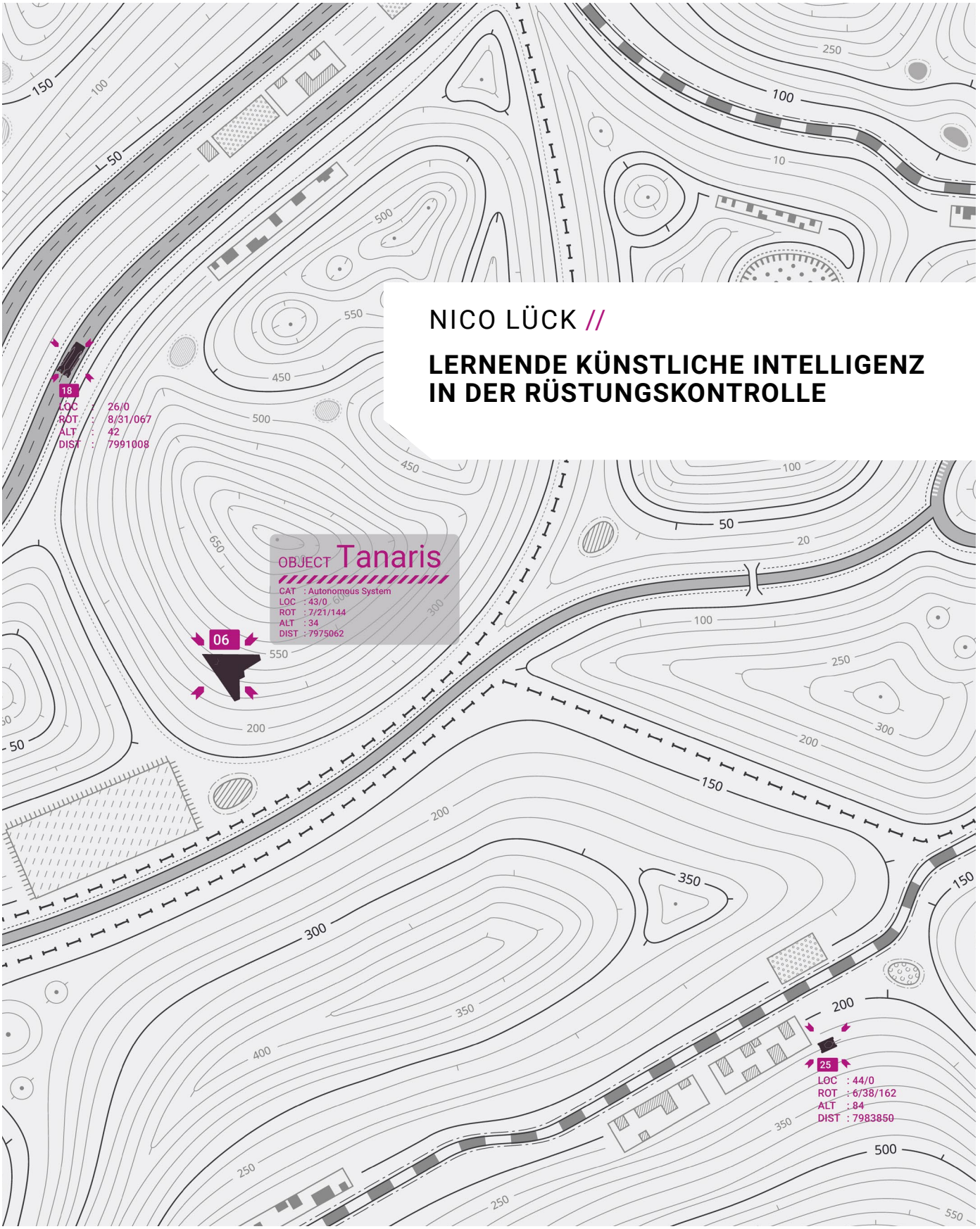
Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
Hessische Stiftung Friedens- und Konfliktforschung (HSFK)

4/2019

PRIF REPORT

PEACE RESEARCH INSTITUTE FRANKFURT / LEIBNIZ-INSTITUT HESSISCHE STIFTUNG FRIEDENS- UND KONFLIKTFORSCHUNG



NICO LÜCK //

LERNENDE KÜNSTLICHE INTELLIGENZ IN DER RÜSTUNGSKONTROLLE

PRIF HSFK
Peace Research Institute Frankfurt
Leibniz-Institut Hessische Stiftung Friedens- und Konfliktforschung

PRIF Report 4/2019

LERNENDE KÜNSTLICHE INTELLIGENZ IN DER RÜSTUNGSKONTROLLE

NICO LÜCK //

Leibniz Gemeinschaft

IMPRESSUM

LEIBNIZ-INSTITUT HESSISCHE STIFTUNG FRIEDENS- UND KONFLIKTFORSCHUNG (HSFK)
PEACE RESEARCH INSTITUTE FRANKFURT (PRIF)

Coverbild:
Nico Lück, eigene Abbildung.

Adresse:
Leibniz-Institut Hessische Stiftung Friedens- und Konfliktforschung (HSFK)
Baseler Straße 27–31
60329 Frankfurt am Main
Telefon: +49 69 95 91 04-0
E-Mail: info@hsfk.de
https://www.hsfk.de

ISBN: 978-3-946459-46-0

Das Thema Künstliche Intelligenz (KI) ist in aller Munde und gilt als eines der zentralen Themen der Zukunft. Denn es ist absehbar, dass die Leistungsfähigkeit von KI-Systemen außergewöhnliche Fortschritte in verschiedensten Anwendungen – Ressourcenoptimierung, Prognose, Objekterkennung, Mensch-Computer-Interaktion oder Steuerung robotischer Systeme – ermöglichen wird. Dies gilt umso mehr, wenn die KI die Fähigkeit zum sogenannten *maschinellen Lernen* besitzt, also die Fähigkeit, auf Grundlage von Beobachtungen oder eines vorgegebenen Datensatzes eigene Regeln zu entwerfen. Diese Fähigkeit ermöglicht eine nie zuvor dagewesene Unabhängigkeit von den Antizipationsfähigkeiten menschlicher Programmierer.

Auch in der Rüstung spielen diese Entwicklungen schon jetzt eine bedeutende Rolle: Klassische KI wird bereits in Kampfflugzeugen, Drohnen, Geschütztürmen oder humanoiden Robotern als Steuerungs- und Assistenzeinheit eingesetzt, etwa zur Navigation und Zielerkennung. Lernende KI wird aktuell in der Entwicklung neuer Waffensysteme getestet oder in Prototypen integriert.

Diese Trends machen Künstliche Intelligenz zu einem wichtigen Thema der Rüstungskontrolle, und zwar in doppelter Hinsicht. Als *Gegenstand* der Rüstungskontrolle entzieht sich KI den bisherigen, traditionellen Ansatzpunkten, da sie weder physische Eigenschaften oder Fähigkeiten noch transparente Funktionsweisen besitzt, auf denen aktuelle Methoden und Verfahren zur quantitativen wie qualitativen Beschränkung in der Rüstung basieren. Auf der anderen Seite gibt KI der Rüstungskontrolle aber auch neue *Werkzeuge* an die Hand. So ist es denkbar, dass die Verifikation bestehender und neuer Rüstungskontrollverträge, also die Überprüfung ihrer Einhaltung, in erheblichem Maß von KI als technischem Hilfsmittel profitieren könnte, etwa durch eine höhere Präzision und Geschwindigkeit bei der Sammlung, Verarbeitung und Analyse von Daten.

Aus sicherheitspolitischer Sicht bietet KI also Risiken und Chancen gleichermaßen, und der abzusehende verstärkte Einsatz von maschinellem Lernen wird beides noch potenzieren. Das *Risiko* besteht darin, dass KI als Kernelement zukünftiger autonomer Waffensysteme durch die Rüstungskontrolle limitiert werden muss, die Rüstungskontrolle jedoch zugleich keine adäquaten technischen Möglichkeiten besitzt. Die traditionellen Ansatzpunkte sind an dieser Stelle erschöpft und es bleiben die Möglichkeiten, die KI während der Entwicklung oder des Einsatzes zu überwachen und zu limitieren. Sollte KI unkontrolliert eingesetzt werden, so gefährdet sie die strategische Stabilität, indem sie den deeskalierenden menschlichen Faktor minimiert, ein technologisches Wettrüsten fördert und sich unkontrolliert verbreiten lässt. Dem steht das enorme *Potenzial* gegenüber, welches selbstlernende Verfahren für die Verifikation von Rüstungskontrollverträgen bergen. Die Identifikation von Objekten, Phänomenen und zeitlichen Veränderungen auf Satellitenbildern, auf Videoaufnahmen oder in Daten von elektromagnetischen, seismischen oder akustischen Sensoren wird durch den Einsatz von maschinellem Lernen nachweislich verbessert. Die weitaus präzisere und umfangreichere Informationsverarbeitung könnte die Transparenz erhöhen, Akteure von einem Vertragsbruch abschrecken oder als Beweis zur Vertragskonformität genutzt werden.

Der Report kommt zu dem Schluss, dass speziell lernende KI sowohl Teil des Problems als auch Teil der Lösung ist. Aktuelle und zukünftige Anwendungsbeispiele zeigen, dass KI das Kernelement moderner Waffensysteme darstellt und damit selbst Gegenstand von Rüstungskontrolle sein soll-

te – speziell da die Fähigkeiten, aber auch die Gefahren, der entsprechenden Waffensysteme durch lernende KI noch einmal potenziert würden. Als Kontrollgegenstand entzieht sie sich jedoch vielen Ansatzpunkten zur qualitativen oder quantitativen Beschränkung. Damit erhöht sie die Relevanz alternativer Methoden zur gesamtmilitärischen Transparenz und Vertrauensbildung. Genau in diesen Methoden kann sie wiederum als verifizierendes Instrument eingesetzt werden. Durch eine präzise und umfangreiche Informationsverarbeitung kann sie erhöhte Transparenz schaffen, die Einhaltung von Verträgen verifizieren und damit Vertrauen zwischen den Parteien stärken. Die Entwicklungen in den beiden Anwendungsmöglichkeiten von KI und der geringeren technischen Komplexität in Verifikationsmaßnahmen zeigen, dass zeitige Handlungen das positive Potential unterstützen können. Denn KI kann bereits jetzt der Rüstungskontrolle zu neuen Kapazitäten verhelfen, bevor diese der neuen Herausforderung KI-gesteuerter Waffensysteme gegenübersteht.

INHALT

1. EINLEITUNG

Der unkontrollierte Einsatz von Künstlicher Intelligenz (KI) in Waffensystemen ist ein Risiko für die Menschheit, das mit der zunehmenden Leistungsfähigkeit der Technologie steigt. Das Risiko zeigt sich, wenn KI absichtlich mit zerstörerischen Aufgaben beauftragt wird oder wenn sie einen solchen Weg eigenständig wählt, um ein vorgegebenes Ziel zu erreichen. Andererseits birgt die Entwicklung von KI jedoch auch positive Chancen für die Menschheit. Durch die außergewöhnliche Leistungsfähigkeit von KI in der Informationsverarbeitung kann sie Muster in scheinbar unstrukturierten Datensätzen erkennen und interpretieren, sowie basierend auf vorgegebenem oder erlerntem Wissen Probleme lösen, Handlungen planen oder Erkenntnisse schlussfolgern. In der Praxis wird KI eingesetzt, um beispielsweise Ressourcennutzung zu optimieren, zeitliche Entwicklungen zu prognostizieren, Objekte in Bildaufnahmen (wieder-)zuerkennen, mit Menschen zu kommunizieren oder robotische Systeme zu steuern. Das Spannungsfeld zwischen Risiken und Chancen des Einsatzes von KI ist daher der grundlegende Gedanke des vorliegenden Reports.

Die Ambivalenz zwischen positiven und negativen Folgen eines KI-Einsatzes findet sich auch in der Rüstungskontrolle: Das übergeordnete Ziel der Rüstungskontrolle ist es, strategische Stabilität zwischen zwei oder mehreren Staaten herzustellen und so die Kriegswahrscheinlichkeit zu senken (Croft 1996: 91–92). Zu diesem Zweck wird die Erforschung, die Herstellung, der Besitz, die Nutzung und die Verbreitung von Rüstungsgütern reguliert (Goldblat 2002: 3).

Wie mit jeder rüstungstechnologischen Innovation wollen Regierungen durch den Einsatz von KI in Waffensystemen eine technologische Überlegenheit und damit einen strategischen Vorteil gegenüber anderen Staaten erlangen oder ausgleichen. Sollte folglich der Einsatz von KI in Waffensystemen reguliert werden, um strategische Stabilität zu verbessern, würde dies die Rüstungskontrolle neu herausfordern. Als Kontrollgegenstand befände sich KI damit in einer Reihe mit Minen, Munition, Kleinwaffen, konventionellen Waffen, Massenvernichtungswaffen und Trägersystemen. Hierbei handelt es sich jedoch um physische Gegenstände, die durch technische, geographische oder anwendungsbezogene Eigenschaften reguliert werden. Soll dies ebenso mit KI geschehen, müssen immanente Eigenschaften von KI identifiziert werden, die eine Beschränkung und Überwachung ermöglichen. Sollte KI hingegen nicht reguliert werden, so können bisher unbekannte Folgen für die strategische Stabilität entstehen.

Diese Gefährdung der Stabilität droht in Zeiten, in denen die Rüstungskontrolle sich bereits in einer Krise durch militär-technischen Fortschritt, Vertragsbrüche und fehlenden politischen Willen befindet (Arbatov 2015; Schmidt 2017). Doch KI bietet gleichzeitig auch neue Möglichkeiten für die Kontrolle von konventionellen Waffen und Massenvernichtungswaffen und kann damit helfen, die Einhaltung bestehender und neuer Rüstungskontrollverträge zu verifizieren. Um das Vertrauen in die Kontrollverträge und zwischen den Staaten zu verbessern, können Überwachungen aus der Ferne und Vor-Ort-Inspektionen eine Transparenz schaffen, die die Konformität eines Staates mit dem Vertrag bestätigen hilft. Zu diesem Zweck kommen meist technische Hilfsmittel zum Einsatz, die Informationen sammeln, verarbeiten und analysieren (Goldblat 2002: 310). Das Potential von Verifikationsmaßnahmen ist mit neuen Technologien gestiegen, da Satelliten, Sensorik und andere Über-

wachungstechniken die Informationslage verbessern (Pilat 2002: 81). An dieser Stelle kann KI als Multiplikator dienen, da sie diese Informationen – vor allem große Datenmengen – mit höherer Präzision und Geschwindigkeit als herkömmliche Methoden analysieren kann.

Der Report erläutert zunächst den Untersuchungsgegenstand „lernende KI" sowie die aktuellen Herausforderungen bei der Entwicklung von Lernfähigkeit. Dann geht er der Frage nach, wo der Einsatz von lernender KI in Waffensystemen droht (Kapitel 3) und wo lernende KI neue Chancen für Rüstungskontrolle und Verifikation eröffnet (Kapitel 4). Der Report schließt in Kapitel 5 mit der Schlussfolgerung, dass sich lernende KI in Waffensystemen vielen traditionellen Rüstungskontrollansätzen entzieht und gleichzeitig KI in Verifikationsmaßnahmen grundlegende Transparenz verbessern und Vertrauen zwischen den Parteien stärken kann.

2. LERNENDE KÜNSTLICHE INTELLIGENZ

Wird versucht, die Frage zu beantworten, was unter KI zu verstehen ist, muss festgestellt werden, dass es viele Ansätze, aber keine allgemein anerkannte Definition gibt. „There are about as many definitions of AI as researchers developing the technology" (McCloskey 2017).

Doch die meisten Definitionen nennen übereinstimmend zwei Kerneigenschaften: 1) die Lösung hoch komplexer Aufgaben und 2) die Anpassungsfähigkeit gegenüber der Umwelt.[1]

KI-Systeme, die in diesen zwei Kerneigenschaften besonders gute Leistungen erreichen, nutzen zumeist die Fähigkeit des Lernens. Die Lernfähigkeit sticht damit unter den oft attestierten Fähigkeiten von KI – Wahrnehmung, Wissensrepräsentation, Problemlösung, Planung und Schlussfolgerung – heraus. Begünstigt wird sie durch die immer noch anhaltende Leistungssteigerung der Hardware, speziell der Rechenleistung der Prozessoren, und die Verfügbarkeit von großen Datensätzen als Lerngrundlage.

In der Vergangenheit konnten Computerprogramme nur Sachverhalte interpretieren, zu denen der Programmierer vorgeschriebene Regeln, d.h. bedingte Wenn-Dann-Beziehungen, festgelegt hatte. Wenn zukünftige Situationen oder zeitliche Veränderungen nicht vom menschlichen Programmierer antizipiert werden können oder wenn der Programmierer die Lösung selbst nicht kennt, dann hilft der Einsatz von maschinellem Lernen (Russell/Norvig 2010: 693). Darunter werden Systeme verstanden, die eigene Regeln durch das Erkennen von Mustern in Datensätzen generieren. Mit dieser Innovation kann KI Probleme der komplexen wirklichen Welt lösen, ohne auf vom Programmierer vorgegebene Lösungswege angewiesen zu sein (Goodfellow 2016: 3). Auf solch lernender KI wird der Fokus dieses Reports liegen, denn erst die Lernfähigkeit ermöglicht herausragende Leistungen in den zwei genannten Kerneigenschaften von KI – Lösungs- und Anpassungsfähigkeit.

---

1 Sammlungen von Definition zu KI finden sich in Artificial General Intelligence Sentinel Initiative (2017); Legg/Hutter (2007).

Zusätzlich zum hier gewählten Fokus auf die spezifische Lernfähigkeit schränkt dieser Report die zu untersuchende KI auch anhand ihrer allgemeinen Zielsetzung ein: Ist das Ziel eine generell intelligente Maschine oder eine Maschine, die nur in einer spezifischen Disziplin als „intelligent" gilt? Im Kern kann jede Definition in diese Zweiteilung eingeordnet werden. Die Vision einer generell intelligenten Maschine zieht zwar das öffentliche Interesse auf sich, jedoch existiert (noch) kein System, welches diese Anforderungen erfüllt. Da Experten die Entwicklungszeit einer sogenannten generellen KI auf mindestens 50 Jahre schätzen (Müller/Bostrom 2016: 559), ist solch eine Definition kein praktikabler Ansatzpunkt für die Untersuchung aktueller und zeitnaher Anwendungsgebiete. Stattdessen werden in diesem Report Beispiele von KI betrachtet, die ausschließlich für eine spezifische Disziplin oder Aufgabe entwickelt sowie optimiert werden. Für diesen bereits existierenden Typus von KI dient der Mensch nicht als Vorbild, sondern wird höchstens als Leistungsvergleich genutzt. Das Ziel der sogenannten anwendungsspezifischen KI – auch *„narrow AI"* (Franklin 2014: 16) oder *„weak AI"* (Searle 1980: 417) genannt – ist es, in mindestens einer Disziplin qualitativ hochwertige Ergebnisse bzw. „intelligente" Leistungen zu erzielen.

Durch die Einschränkungen auf lernende anwendungsspezifische KI zielt die Betrachtung der Empirie in diesem Report vorrangig auf neue, innovative KI-Programme und schließt sowohl herkömmliche Computerprogramme als auch futuristische KI-Konzepte aus. In Bezug auf die genutzten Begrifflichkeiten bleibt festzuhalten, dass der Begriff „maschinelles Lernen" die Fähigkeit bezeichnet, die den Untersuchungsgegenstand „lernende KI" charakterisiert. In diesem Report wird bewusst zwischen den Nennungen von „KI" und „lernender KI" differenziert. Wenn nur von „KI" die Rede ist, dann sind die Aussagen auch für KI ohne Lernfähigkeit gültig oder der Funktionsweise eines empirischen Beispiels ist nicht mit Sicherheit die Lernfähigkeit zu entnehmen.

2.1 TREIBER DER KI-REVOLUTION: MASCHINELLES LERNEN

Viele Handlungen von Menschen oder Tieren sind für Computer hoch komplex. Die Verarbeitung von Sprache oder Bildern und anschließende Koordination von Aktionen hat eine solch hohe Komplexität, dass Menschen den Verarbeitungsprozess selbst nicht ausreichend verstehen, um ihn auf ein Programm übertragen zu können. Damit diese hoch komplexen Handlungen auch von KI bewältigt werden können, wird ihr die Fähigkeit zu lernen hinzugefügt. Lernende KI erkennt Lösungsmuster und transferiert diese auf andere Herausforderungen (Shalev-Shwartz/Ben-David 2014: 21–22). Dies ist notwendig, da sich die Anforderungen an den auszuführenden Prozess je nach Zeit, Benutzer oder anderen Parametern verändern können.

Zwei Formen des maschinellen Lernens lassen sich unterscheiden: KI kann allein auf Grundlage von gegebenen Daten (*unbeaufsichtigt*) oder mit Hilfe von weiteren Inputs eines Lehrers lernen (*beaufsichtigt* oder *bestärkend*).

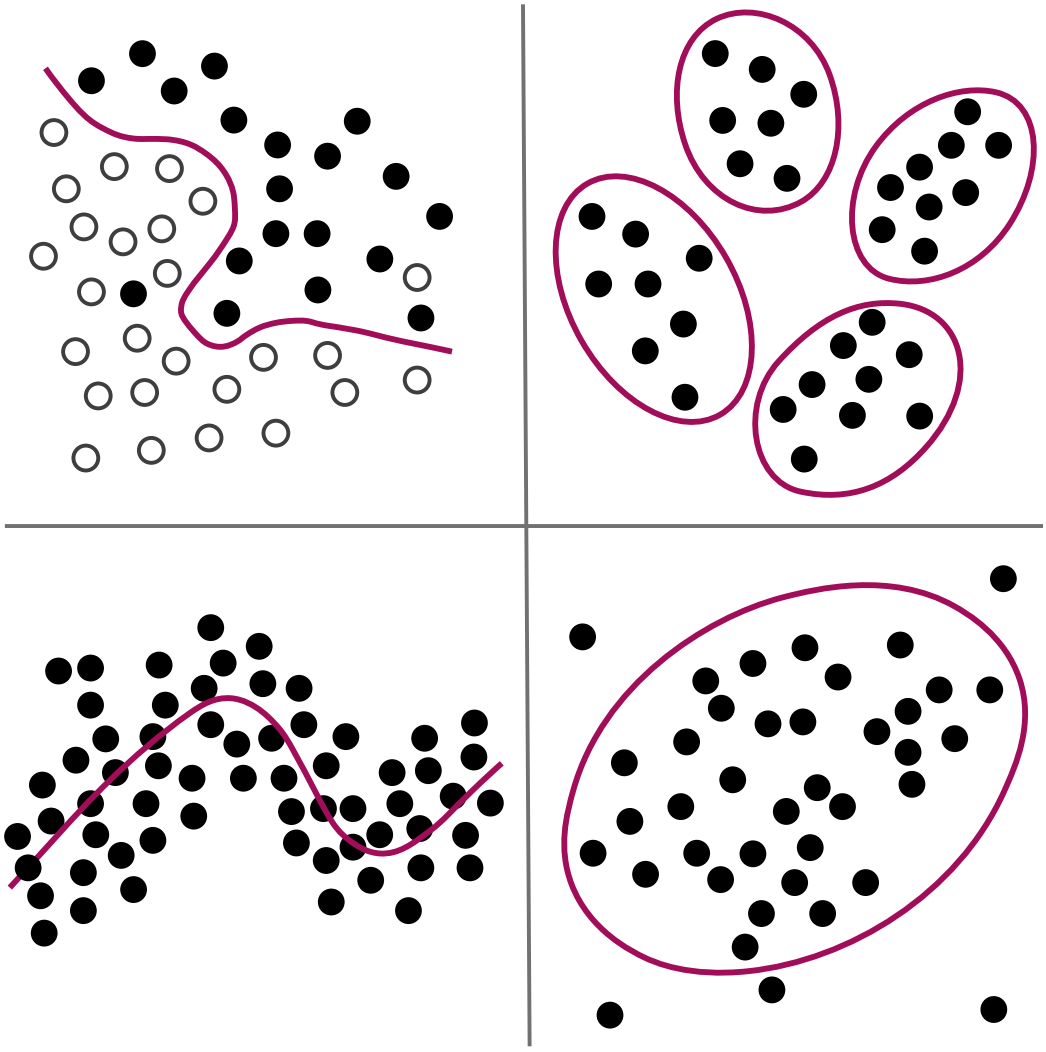


Abb. 1: Klassifikation (oben links), Cluster-Bildung (oben rechts), Regression (unten links), Erkennung von Anomalien (unten rechts). Eigene Darstellung.

Unbeaufsichtigtes Lernen findet Anwendung, wenn lernende KI aus unbekannten Daten Cluster bilden soll. Die KI findet selbstständig Merkmale, die Gemeinsamkeiten und Unterschiede darstellen. Sie erkennt beispielsweise, dass Satellitenbilder nach sichtbarem Land, Meer oder Wolken gruppiert werden können. Beim beaufsichtigten bzw. bestärkenden Lernen hingegen werden die relevanten Merkmale der Daten vordefiniert. Diese Vorgehensweise ermöglicht eine Klassifikation (z.B. Zuordnung von Subjekten in Bildern), Regression (z.B. Vorhersage von Umweltphänomenen) oder Erkennung von Anomalien (z.B. Erkennung unnatürlicher Eruptionen) innerhalb der Daten (Russell/Norvig 2010: 694–697).

Die Formen des maschinellen Lernens können durch zahlreiche Methoden wie *Naive Bayes*, *Support Vector Machine*[2] oder diversen Varianten von Entscheidungsbäumen umgesetzt werden. Doch besonders die aktuelle Methode *Deep Learning* (auch bekannt als *Deep Neural Network*) verantwortet einen Großteil des aktuellen Fortschritts in der KI-Forschung. Ihre hohe Anpassungs- und Leistungsfähigkeit bei hoch komplexen Aufgaben liegt in der Orientierung am menschlichen Gehirn begründet, indem sie dessen Netz aus Neuronen nachbildet. Doch bisher kommt auch sie nicht annähernd an die Komplexität des menschlichen Gehirns heran (Hawkins/Blakeslee 2004: 25–27). Jedoch ist der erreichte Leistungsgewinn für spezifische Anwendungen wie der Bilderkennung bereits beträchtlich. Während ein herkömmliches Programm daran scheitert, aus einer Ansammlung von verschiedenen Pixeln eines Bildes eine Bedeutung zu extrahieren, unterteilt die *Deep-Learning*-Methode den Prozess in viele Teilschritte. Die erste Ebene vergleicht die Helligkeit der Pixel, um Kanten im Bild zu identifizieren. Die zweite Ebene sucht aufgrund der Ergebnisse der ersten Ebene nach Ecken und Konturen. Die dritte Ebene wiederum kann darauf aufbauend spezifische Objekte (z.B. Nase, Ohren, Beine) durch die Kombination aus Ecken und Konturen identifizieren. Schlussendlich kann der Computer das Bild durch die verschiedenen erkannten Objekte interpretieren (Goodfellow 2016: 6). Dieser Prozess kann nur durchgeführt werden, nachdem der Computer vorher gelernt hat, welche Objekte ein Bild beispielsweise mit einem Auto, einer Person oder einem Tier beinhaltet. Im Lernprozess ist der „Output Layer" gegeben und die Ebenen werden in umgekehrter Reihenfolge abgearbeitet. Die lernende KI assoziiert zu dem gegebenen Ergebnis die erkannten Objekte der dritten Ebene, zu den Objekten die Konturen der zweiten Ebene sowie den Konturen die Helligkeit der Pixel der ersten Ebene[3]. Es ist äußerst wichtig zu verstehen, dass die Ergebnisse, zu denen das *Deep Learning*-Programm gelangt, nur Wahrscheinlichkeiten widerspiegeln. Es entscheidet sich nicht für eine Option und begründet diese, sondern es gibt beispielweise an, es handle sich mit 9%iger Wahrscheinlichkeit um ein Auto, mit 72% um eine Person oder mit 19% um ein Tier. Wie diese Ausgaben weiter genutzt werden, liegt in der Hand des Programmierers.

Da Maschinen nur die formalisierten Interdependenzen der Variablen verstehen und nicht ihren Inhalt, wird maschinelles Lernen von manchen Wissenschaftlern und Kritikern als automatisierte Statistik bezeichnet (Danks 2014: 159f). Allerdings basieren die Lernmodelle selbst oder das übergeordnete Programm auf einem regelbasierten System. Dadurch beinhalten sie – im Gegensatz zu Statistik – immer bewusste oder unbewusste normative Vorstellungen, die durch die Programmierer über Entscheidungen im Programmcode oder die Auswahl der Lerndaten eingebracht werden (Algorithm Watch 2017: 3).

Zu bedenken ist, dass das Potential maschinellen Lernens noch lange nicht ausgereizt ist. Zahlreiche unter den Begriff fallende Methoden wurden bereits und werden noch entwickelt (Farrelly 2016). Neben dem Erreichen einer gesteigerten Komplexität sind zukünftige Meilensteine in folgenden Bereichen zu erwarten: kontinuierliches Lernen über ein begrenztes Set aus Eingabedaten hinaus,

2 Diese Methoden basieren auf mathematischen Verfahren, die zur Klassifizierung von Objekten unterschiedliche Variablen in Funktionen maximieren, beispielweise mit dem Ziel einer größtmöglichen Differenz zwischen den Objekten oder der niedrigsten Klassifizierungskosten bei einem vordefinierten Maß.

3 Eine ausführlichere Erklärung und hervorragende Visualisierung des Prozesses findet sich in Zeiler und Fergus (2013).

Transfer des erlernten Wissens auf andere Aufgaben, die eigenständige Generierung der Eingabedaten (Morisse 2017) und Lernen durch Beobachtung anderer Maschinen (Li et al. 2016).

2.2 HERAUSFORDERUNGEN BEI DER ENTWICKLUNG UND NUTZUNG EINER LERNENDEN KI

Soll lernende KI in einer realen Anwendung genutzt werden, so stellen sich drei Herausforderungen: (1) Fehler oder Ungenauigkeiten im System oder Lernprozess, (2) schwere Nachvollziehbarkeit der Entscheidungen bei mangelnden Vorkehrungen und (3) Manipulierbarkeit der Ergebnisse mit präparierten Eingabedaten.

Die erste hier dargestellte Herausforderung maschinellen Lernens ist es, eine Balance zwischen zwei Faktoren herzustellen, nämlich zwischen der Komplexität des erlernten Modells, das mit zunehmender Präzision gleichzeitig an Komplexität gewinnt, auf der einen Seite und der Generalisierbarkeit des erlernten Modells auf neue Daten auf der anderen Seite (Danks 2014: 155f). Die beiden Faktoren stehen unter einer dilemmatischen Wechselwirkung: Wird das Modell zu komplex und bildet exakt die Trainingsdaten – also die Daten, die die Realität beschreiben sollen – ab, so kann das Modell nicht mehr generalisieren, also nicht auf neue, bisher unbekannte Daten angewandt werden (*Überanpassung*). Bildet das Modell die Trainingsdaten zu ungenau ab, so kann zwar generalisiert werden, aber die Realität wird nicht korrekt modelliert (*Unteranpassung*). Die Generalisierung ist infolgedessen anfällig für Trugschlüsse. Es können schädliche Feedbackschleifen entstehen, in denen die lernende KI eine eigene Logik entwirft, aus der sie ohne korrigierendes Feedback nicht ausbrechen kann. In der Praxis werden diese Generalisierungsfehler nicht immer entdeckt und daher in reale Anwendungen übernommen. Die Akteure, die von diesem falschen Modell der KI abhängig sind, müssen den von der KI erstellten Regeln folgen, um Erfolg zu haben. Dies verfälscht wiederum das mögliche Feedback an das System. Diese Problematik zeigt sich beispielsweise bei der Vorhersage von Straftaten: Polizeibeamte fahren gesondert in Bezirken Streife, die das System zur aktuellen Tageszeit als besonders kritisch bewertet. Durch die erhöhte Polizeipräsenz werden mehr Straftaten entdeckt. Die Vorhersage des Systems wird bestätigt und gleichzeitig werden die protokollierten Vorfälle in die Datengrundlage für weitere Vorhersagen aufgenommen. Solche Feedbackschleifen konnten auch in der Gesichtserkennung, Lehrerbewertung sowie Kredit- und Versicherungsvergabe nachgewiesen werden (O'Neil 2016).

Als zweite Herausforderung ist besonders bemerkenswert, dass maschinelles Lernen, insbesondere *Deep Learning*, aufgrund inhärenter Charakteristika von vielen Forschern als „Black Box" bezeichnet wird (Knight 2017; Ribeiro et al. 2016). Denn die selbst erlernten Regeln sind in ihrer Gesamtheit ein mathematisches Modell aus Tausenden oder sogar Millionen von Parametern. Ein Entscheidungsweg in diesem komplexen Modell, beispielsweise die Zuordnung eines kleinen Bildausschnittes zu einer Kategorie, kann durch einen Menschen nicht logisch nachvollzogen werden. Selbst die Entwickler können nicht immer nachvollziehen, wie die KI zu ihrer Entscheidung kommt. In vielen Fällen wird diese Ungewissheit hingenommen oder dem System schlicht vertraut. Doch dies wird mit der Verbreitung der Methode *Deep Learning* ein zunehmendes Risiko. Andere Methoden

sind zwar in höherem Maße nachvollziehbar, allerdings korreliert die Nachvollziehbarkeit negativ mit der Genauigkeit. Lernt ein Programm beispielsweise mithilfe eines Entscheidungsbaums mit binären Zweigstellen, so sind die Ergebnisse zwar nachvollziehbar, jedoch nicht so genau wie die simultane, prozentuale Gewichtung mehrere Merkmale auf hunderten Abstraktionsebenen eines *Deep Learning*-Modells (Gunning 2016: 4). Im Falle von *Deep Learning*-Systemen wird versucht, die Nachvollziehbarkeit zu erhöhen („explainability"), indem nicht nur eine Wahrscheinlichkeit für eine Option angegeben wird, sondern eine Begründung, warum diese Option die wahrscheinlichste ist. Anstatt einer Ausgabe von „Das Objekt ist zu 93% eine Katze" sollte die lernende KI beispielsweise angeben: „Das Objekt ist zu 93% eine Katze, weil es Fell, Tatzen und Krallen besitzt". Forschungen über die Verknüpfung von Mustererkennung und sprachlicher Beschreibung des Musters zeigen, dass zusätzliche KI-Anwendungen solch eine Nachvollziehbarkeit ermöglichen können (Park et al. 2017). Auch wenn einige Ansätze vielversprechend klingen, ist dieser Forschungszweig[4], zumindest im öffentlich einsehbaren Bereich, noch jung und klein. Die US-militärische Forschung hingegen hat dem Thema nachvollziehbarer KI einen eigenen Programmbereich gewidmet (Gunning 2016).

Die dritte Herausforderung wurde durch die Erkenntnisse eines Forscherteams von Google, Facebook und diversen Universitäten sichtbar. Das Team fand heraus, dass die Bilderkennung der *Deep-Learning*-Methode eine unerwartet hohe Fehlerrate aufweist und damit falsch klassifiziert, wenn dem Bild für das menschliche Auge unsichtbare Informationen hinzugefügt wurden (Szegedy et al. 2014). Diese Ergebnisse lassen zwei fundamentale Erkenntnisse zu: *Erstens*, selbst KI mit herausragender Leistungsfähigkeit lernt nicht das wahre Konzept, das den Bildern zugrunde liegt. Stattdessen konstruiert sie aus statistischen Zusammenhängen ein Modell, welches zwar natürlich auftauchende Daten einbezieht, jedoch fundamentale Schwächen zeigt, wenn es mit einer sehr unnatürlichen bzw. unwahrscheinlichen Datenverteilung konfrontiert wird (Goodfellow et al. 2015: 2). *Zweitens* stellt die Möglichkeit, maschinelles Lernen mit Eingabedaten zu manipulieren, eine große Sicherheitslücke dar. Ohne eine Lösung besteht ein konstantes Risiko, dass Daten, die die lernende KI analysiert, absichtlich so geändert wurden, dass sie die Daten falsch interpretiert. Darüber hinaus konnten Wissenschaftlerinnen und Wissenschaftler zeigen, dass ein konfliktäres Bild (*adversarial example*) die manipulierenden Eigenschaften beibehält, wenn es ausgedruckt und mit irgendeiner Kamera erneut fotografiert wird. So kann ein Bild von einer Waschmaschine für die lernende KI wie ein Safe aussehen (Kurakin et al. 2017). In einer weiteren Studie konnte gezeigt werden, dass Gesichtserkennungssysteme mit speziellen Brillen so beeinflusst werden konnten, dass die aufgenommenen Personen als andere Personen aus der Datenbank identifiziert wurden. Die Autoren warnen davor, dass solche Methoden zukünftig in kriminellen Akten genutzt werden könnten (Sharif et al. 2016). Diese optische Illusion für Maschinen birgt enorme Risiken für praktische Anwendungen von lernender KI. Würden Angreifer ein autonom fahrendes Auto manipulieren wollen, so können sie Straßenschilder so verändern, dass die KI, die Verkehrszeichen erkennt und interpretiert, anstatt eines Stoppschildes ein Vorfahrtsschild als Signal weitergibt. Dass dies nicht nur ein hypothetisches Szenario ist, bewiesen die Forscher um Nicolas Papernot, indem sie das beschriebene Szenario rekonstruierten (Papernot et al. 2017). Diese Fälle verdeutlichen, dass die Manipulationen nicht nur durch digitale Änderungen an der Datei, sondern durch rein visuelle Änderungen möglich sind, da auch Gegenstände in der realen Welt

4 Diese Forschung wird mit dem Begriff *Explainable Artificial Intelligence* (XAI) bezeichnet.

manipuliert werden können. Zudem wurden die Lernmodelle der Waschmaschinen-, Gesichts- und Verkehrsschilderkennung ordnungsgemäß trainiert, jedoch wurden die Anwendungsdaten – also die Daten, die analysiert werden sollen – manipuliert.

3. LERNENDE KÜNSTLICHE INTELLIGENZ IN WAFFENSYSTEMEN

Unter dem Dach der UN findet aktuell in Genf eine Debatte um das Verbot von tödlichen autonomen Waffensystemen statt (Boulanin/Verbruggen 2017). Auch in der Zivilgesellschaft werden Aufrufe von prominenten Wissenschaftlerinnen und Wissenschaftlern sowie Organisationen unterstützt, autonome Waffensysteme zu verbieten (Sauer 2016; Future of Life Institute 2015; Human Rights Watch 2012). Diese Debatten und Aufrufe thematisieren zwar auch den Einsatz von KI in Waffensystemen, doch fokussieren sie sich auf die Eigenschaft der Autonomie. Wie in diesem Kapitel gezeigt werden wird, stellt Autonomie als Eigenschaft von KI-gesteuerten Systemen keinen geeigneten Ansatzpunkt für die Rüstungskontrolle dar. Da auch andere Ansatzpunkte der traditionellen Rüstungskontrolle (physische Eigenschaften oder Fähigkeiten sowie innere Funktionsweise) im Fall von KI keine verlässlichen Nachweise bieten, werden alternative Kontrollmethoden in der Entwicklung und während des Einsatzes betrachtet. In Anbetracht eines möglichen Einsatzes von KI in Waffensystemen werden destabilisierende Folgen wie beschleunigte Prozesse, Mangel an deeskalierenden Handlungsmöglichkeiten, technologisches Wettrüsten oder unkontrollierte Verbreitung diskutiert.

3.1 KI ALS KERNELEMENT MODERNER WAFFENSYSTEME UND LERNFÄHIGKEIT ALS MULTIPLIKATOR

So wie bei herkömmlichen Computern existiert auch bei Waffensystemen eine Trennung zwischen Hard- und Software. Diese Unterscheidung spiegelt sich in den Begriffen Robotik und KI wider. Entwicklungen in der Robotik verbessern die physische Handlungsfähigkeit und die Feuerkraft von Waffensystemen. Doch deren Verbesserungspotential ist durch naturgesetzliche physikalische Restriktionen limitiert. Die Entwicklung von lernender KI hingegen potenziert die softwarebedingten Handlungsfähigkeiten von Waffensystemen um ein Vielfaches. Denn erst lernende KI verbessert zunehmend die Fähigkeit von Waffensystemen, in komplexen Umwelten zu agieren.

Auch wenn noch keine allgemeingültige Definition autonomer Waffensysteme existiert, findet sich anwendungsspezifische KI bereits in Waffensystemen als Steuerungs- oder Assistenzeinheit. Die Lernfähigkeit ist den folgenden Beispielen nicht mit Sicherheit zu entnehmen, jedoch wird – je nach System – KI zur Navigation, Zielerkennung und -identifikation sowie Angriffsplanung und -ausführung genutzt. Viele Anwendungen finden sich im digitalen Raum, im Luftraum und in statischen Verteidigungssystemen, da die Umgebung, in der die KI in diesen Verwendungen operieren muss, weniger komplex als im Boden- oder Häuserkampf ist. Die Umwelt nimmt an Komplexität zu, wenn die Anzahl an (unbekannten) Herausforderungen steigt: z.B. Navigation auf uneinheitlichem Grund, Hindernisse aller Art und Interaktion mit fremden Objekten.

Als Assistenzsystem für Kampfpiloten wird eine KI z.B. darauf trainiert, Ziele anhand des Radars zu erkennen, um Fehleinschätzungen des Piloten zu vermeiden oder aus größerer Entfernung deutlich jenseits der Sichtweite des Piloten eine Abschussentscheidung zu treffen (Keller 2015). Mehr Aufgaben übernimmt KI in der Drohne *Taranis, welche der* britische Hersteller *BAE Systems derzeit entwickelt*. Diese besitzt zusätzlich zur manuellen Fernsteuerung und automatischen Flugnavigation einen Modus, in welchem sie eigenständig eine Route entwirft und nach Zielen sucht, bis sie das Missionsziel erreicht (Stevenson 2016). Die bisher nicht in Drohnen verbaute KI *ALPHA* ist in der Lage, den gesamten Flug und die Kampfmanöver zu übernehmen – bis vor Kurzem noch exklusive Domäne menschlicher Piloten: In einer Simulation gegen den erfahrenen Oberst der *US Air Force*, Gene Lee, bewies das System herausragende Stärke, indem es simultan Geschossen auswich, auf mehrere Ziele feuerte, sich an koordinierten Manövern beteiligte und feindliche Taktiken registrierte und davon lernte. Die KI, die auf einem nur 35 USD teuren Computer, einem *Raspberry Pi*, lief, bezeichnete der Oberst anschließend als „the most aggressive, responsive, dynamic and credible AI I've seen to date" (Ernest et al. 2016).

Auch einige statische Verteidigungssysteme – kleine Geschütztürme oder Luftabwehrgeschütze – sind unter den ersten KI-gesteuerten Waffensystemen zu finden, da sie auf keine komplexen Herausforderungen in der Umwelt treffen. Der Geschützturm *Super aEgis II* wurde dafür konzipiert, selbstständig und ohne menschliche Hilfe Ziele zu identifizieren, anzuvisieren, zu verfolgen und schließlich zu feuern. Aufgrund der Befürchtungen der Kunden, das System könne im autonomen Modus Fehler machen, kann nun der Grad der Autonomie individuell eingestellt werden (Parkin 2015). Auch das Luftabwehrgeschütz *Phalanx-CIWS* der US Navy kann diese Handlungen autonom durchführen, um anfliegende Geschosse und Luftfahrzeuge abzuwehren (US Navy 2017).

Die Anwendungsgebiete werden stetig ausgebaut und autonome Nanodrohnen (Daniels 2017), Kriegsschiffe (Courtland 2016) oder humanoide Roboter (Boston Dynamics 2018) befinden sich bereits in Entwicklung. Doch die Entwicklung fokussiert sich nicht nur auf individuelle Systeme, sondern auch auf eine neue Art der Interaktion. Waffensysteme werden in Zukunft auch im Schwarm agieren können. Ein Schwarm besteht aus vielen individuellen Maschinen, die eigenständig, aber auch gemeinsam agieren können. Durch ununterbrochene Kommunikation zwischen den Einheiten koordinieren sie sich selbst (Ben-Ari/Mondada 2018: 251–252). Dies hat unter anderem den Vorteil, dass es keine zentrale Steuereinheit gibt, die ausfallen kann, sondern dass einzelne Defekte oder Abschüsse nur einen geringen Effekt auf die Leistungsfähigkeit des Schwarms haben. KI würde hier also nicht nur die technische Überlegenheit von einzelnen Systemen ermöglichen, sondern die Einsatzmöglichkeit von ganzen Verbänden optimieren können.

Ein bereits im militärischen Betrieb befindliches Waffensystem, das die Fähigkeit des maschinellen Lernens nutzt, ist bisher nicht bekannt. Die beschriebene KI *ALPHA* zur Steuerung eines Kampfflugzeuges sowie eine Bekanntgabe des russische Rüstungskonzerns *Kalaschnikow*, die Lernfähigkeit zu nutzen (Russia Today 2017), scheinen zu zeigen, dass lernende KI in neue Waffensysteme integriert wird. Auch wenn die Funktionsweise aufgrund der militärischen Geheimhaltung unbekannt ist, vermitteln zivile Forschungsprojekte eine Idee, wie lernende KI in Waffensystemen eingesetzt werden könnte. Ein Forschungsteam des Chipherstellers Nvidia, der eigentlich hochspezialisierte Grafikchips

entwickelt, trainierte eine KI dazu, ein Auto zu steuern, ohne irgendwelche Regeln vorzugeben. Als Eingabedaten bekam die KI nur die Bewegungen des Lenkrads und die Aufnahmen einer Kamera in der Front des Autos. Trotz dieser beschränkten Wahrnehmung lernte die KI im Laufe der menschlich gesteuerten Fahrten die Regeln des Straßenverkehrs und konnte anschließend eigenständig fahren (Bojarski et al. 2016). Dies stellt einen großen Unterschied zu herkömmlichen autonomen Fahrsystemen dar, welchen die Interpretation von Verkehrsregeln und Fahrzeugverhalten vorgegeben wird. Die erlernte Fahrweise hat den offensichtlichen Nachteil, dass menschliche Fehler übernommen werden. Sie birgt jedoch auch große Vorteile, da das Programm für den Fahrer unbewusste, intuitive Regeln erlernt und auch auf nicht antizipierte Situationen trainiert wird. Mobile Waffensysteme könnten demnach analog durch beaufsichtigtes Lernen die Fähigkeit der Navigation erlangen. Auch die Zielerfassung und -identifikation kann durch maschinelles Lernen erheblich präzisiert werden, indem die typischen Fähigkeiten der *Deep Learning*-Methode – Wahrnehmung und Klassifikation von Objekten – eingesetzt werden. In (Kampf-)Situationen muss KI angemessen entscheiden und handeln. Die dazu nötigen Schlussfolgerungs- und Planungsfähigkeiten können mithilfe von Simulationen, die von Menschen überwacht werden, trainiert werden. Demnach kann maschinelles Lernen die Leistung aller Fähigkeiten verbessern, die eine KI in Waffensystemen benötigt.

3.2 KONTROLLERSCHWERENDE EIGENSCHAFTEN UND NEUE ANSATZPUNKTE

Wenn KI tatsächlich das Kernelement autonomer Waffensysteme darstellt und Lernfähigkeit die Möglichkeiten dieser Systeme ohne eine benötigte Anpassung der Hardware weiter verbessert, dann lässt sich daraus schließen, dass letztlich nicht die Eigenschaft der Autonomie, sondern die steuernde KI im Fokus von Rüstungskontrolle stehen sollte. Lernende KI, das soll nun gezeigt werden, führt aber zu ganz neuen Problemen für die Rüstungskontrolle.

3.2.1 AUSWECHSELBARES ÄUSSERES DURCH ERHÖHTE HARDWAREKOMPATIBILITÄT

Wie in Kapitel 3.1 thematisiert, kann der Entwicklungsprozess von Hard- und Software getrennt betrachtet werden. Weil die Interaktion der beiden Ebenen in praktischen Anwendungen abgestimmt sein muss, kann jedoch keine pauschale Auswechselbarkeit der Hardware abgeleitet werden. Die Software muss mit der Hardware kommunizieren können. Eine Möglichkeit, die Kompatibilität mit einer Vielzahl von Hardware-Bauteilen zu erhöhen, sind einheitliche Standards und automatische Treiberaktualisierungen. Für komplexere Hardwaresysteme wird in der Robotik sogenannte Middleware eingesetzt. Die Middleware regelt die Heterogenität der Hardware und der Anwendungen durch eine zusätzliche Ebene. Sie erleichtert die Integration neuer Technologien, die Nutzung der Sensordaten und die Austauschbarkeit von Bauteilen. Die Integration einer lernenden KI kann die erzeugte Kompatibilität der Middleware weiter erhöhen, indem sie ihr eine dynamische Anpassung an das System ermöglicht (Bennaceur et al. 2013). Sollten solche Methoden auch in der Entwicklung von Waffensystemen genutzt werden, könnte lernende KI ohne aufwändige Anpassungen in unterschiedlichen Waffensystemen eingesetzt werden. Eine KI könnte gleichermaßen eine Drohne, ein Unterwasser-

fahrzeug, eine Rakete oder andere robotische Waffen steuern. Somit kann das Kernstück neuer Waffentechnologien mit keinem spezifischen Waffensystem exklusiv assoziiert werden.

Ein üblicher Ansatzpunkt der Rüstungskontrolle ist die quantitative Begrenzung des Trägersystems einer Waffe. Wird die KI als Waffe bzw. dessen Multiplikator verstanden, so stellen konventionell bewaffnete Drohnen, Roboter usw. die Trägersysteme dar. Eine Limitierung des Trägersystems würde jedoch nur den Transfer der KI auf ein anderes System bewirken. Navigation, Zielerfassung und Handlung könnte auf allen Systemen in einem ähnlichen Schema stattfinden und damit die Kompatibilität herstellen, die ein Transfer benötigt. KI als zerstörerisches Kernelement von Waffensystemen kann demnach nicht zu einem äußerlich sichtbaren System assoziiert werden und nimmt der Rüstungskontrolle damit einen üblichen Ansatzpunkt.

3.2.2 AUSTAUSCHBARKEIT ÄUSSERLICH SICHTBARER FÄHIGKEITEN DURCH SOFTWAREAKTUALISIERUNGEN UND OFFENE SOFTWAREARCHITEKTUR

Eine typische Eigenschaft von Computerprogrammen sind Aktualisierungen, die Sicherheitslücken schließen, Funktionen hinzufügen oder Komponenten verändern sollen. Mit der zunehmenden Nutzung von softwaregestützten Waffensystemen sind Updates auch dort notwendig. Hoch technisierte Kampfflugzeuge zeigen dies beispielhaft: Die *US Air Force* aktualisierte die Software des F-22 Kampfflugzeugs, um neuere Waffen abfeuern, Ziele besser identifizieren und damit ein weiteres Spektrum an Angriffsmissionen durchführen zu können (Osborn 2017). Auch Waffensysteme, die eine lernende KI nutzen, können demnach mit zusätzlichen Funktionen ausgestattet werden, ohne dass die Hardware verändert wird. Zur weiteren Flexibilisierung dieses Prozesses kann zusätzlich eine offene Softwarearchitektur eingeführt werden. Solch eine Architektur findet sich in Smartphones: So genannte „Apps" sind Applikationen, durch die dem System Komponenten hinzugefügt, entfernt oder aktualisiert werden können, ohne dass das Hauptprogramm verändert wird. In der Rüstung wird dieses System bereits in dem F-35 Kampfflugzeug angewendet. Das israelische Militär importiert dieses Flugzeug und kann es mithilfe der offenen Architektur an die eigenen Ansprüche anpassen, ohne die zentrale Software zu verändern (Adams 2016). Auch andere US-amerikanische Rüstungskonzerne entwickeln solch eine offene Softwarearchitektur für die eigenen Produkte. Würde ein konzernübergreifender Standard gelten, könnten Applikationen unabhängig von Art und Bauweise des Waffensystems flexibel eingesetzt werden (Hagen et al. 2012: 6).

Ein Ansatzpunkt der Rüstungskontrolle zielt auf die Beschränkung der Fähigkeiten einer Waffe. Der *Kernwaffenteststopp-Vertrag* verbietet, sobald er in Kraft tritt, die Durchführung von Kernwaffenexplosionen für zivile oder militärische Zwecke. Die Rüstungskontrolle setzt hier an der Explosionsfähigkeit von Kernwaffen an. Die Fähigkeiten einer KI in Waffensystemen können allerdings flexibel hinzugefügt und entfernt werden. Anhand der sichtbaren Fähigkeiten sind demnach keine Identifizierung und keine Einschränkung von Waffensystemen möglich. Auch bei Inspektionen könnten kritische Funktionen kurzzeitig hinzugefügt oder entfernt werden.

3.2.3 INTRANSPARENTE INNERE FUNKTIONSWEISE DURCH ERSCHWERTES REVERSE ENGINEERING UND MANGELNDE NACHVOLLZIEHBARKEIT DES LERNMODELLS

Soll eine Vereinbarung über die Funktionsweise eines Computerprogramms verifiziert werden, kann der Quellcode des Programms analysiert werden. Liegt nur das fertige System vor, beispielweise eine autonome Drohne, kann der Quellcode jedoch nicht ohne weiteres extrahiert werden. Programme werden typischerweise in einer sogenannten Hochsprache geschrieben, die meistens durch einen Compiler in Maschinensprache umgewandelt wird.[5] Durch diese Umwandlung gehen Metainformationen verloren, die die Umkehrung des Prozesses erschweren. Durch spezielle Programme kann eine aufwändige Rekonstruktion im Rahmen des sogenannten *Reverse Engineerin*g gelingen (Eilam 2005). Dennoch entsteht eine erste Hürde zur Kontrolle und Verifikation von digital gesteuerten Waffensystemen allein durch die grundlegende Architektur von Software. Daher gilt dies bereits für heutige Waffensysteme, auch ohne lernende KI.

In der Anwendung maschinellen Lernens kommt allerdings eine zweite Ebene der Intransparenz hinzu. Wie in Kapitel 2.2 beschrieben, bestimmen inhärente Charakteristika der verschiedenen Lernmethoden die Transparenz der KI. Bisher wenig erforschte Methoden müssten dem Waffensystem hinzugefügt werden, damit es seine Handlungen begründen könnte. Neben einer nachträglichen Erklärung ist auch eine vorherige Determinierung der Handlungen unmöglich, solange das erlernte Modell nicht einsehbar ist.

Ein weiterer üblicher Ansatzpunkt zur Rüstungskontrolle stellt die Funktionsweise beispielweise von Kernwaffen, Antipersonenminen und Streumunition dar. Bei der Kernwaffe wird die Nutzung der Kernenergie für eine Explosion kontrolliert. Antipersonenminen und Streumunition sind verboten, da die Funktionsweise nicht zwischen Kombattanten und Zivilisten unterscheiden kann. Bei lernender KI wird die Funktionsweise durch die zwei beschriebenen Ebenen intransparent. Diese Intransparenz nimmt der Rüstungskontrolle einen weiteren Ansatzpunkt, da lernende KI in Waffensystemen nicht aufgrund der Funktionsweise definiert und begrenzt werden kann.

3.2.4 TECHNISCHE ANSATZPUNKTE FÜR EINE KONTROLLE

Wenn – wie in den vorherigen Kapiteln geschlussfolgert – die Existenz, die Hardware, die Funktionsweise und die Fähigkeiten von lernender KI keine verlässlichen Ansatzpunkte für eine Kontrolle bieten, dann muss dies über die Entwicklung oder den Einsatz geschehen. Bei der Entwicklung setzt das Konzept der präventiven Rüstungskontrolle an. Dort sollen bereits in der Entwicklungs- oder Erprobungsphase militärisch nutzbare Techniken, Stoffe oder Systeme erkannt und verboten bzw. reguliert werden (Altmann 2008):

5 Hochsprache kann auch in Echtzeit interpretiert werden. Eine direkte Interpretation macht das Testen und Ändern des Codes einfacher. Das Kompilieren ist dennoch ein üblicher Schritt, da es die Effizienz des Codes erhöht.

> „Konkret zielt präventive Rüstungskontrolle darauf ab, die entsprechenden Forschungs- und Entwicklungsprozesse zu begrenzen, zu unterbrechen, oder zu beenden und/oder die auf ihrer Umsetzung in Waffen(systeme) aufbauenden militärischen Optionen zu verbieten." (Neuneck/Mutz 2000: 109).

In diesem Zusammenhang wäre ein Register für militärische Forschung und Entwicklung denkbar, welches Rüstungsrisiken frühzeitig erkennen ließe (Müller 2000). Doch dieser Ansatzpunkt erfordert hohe Transparenz im Entwicklungsprozess, die weit leichter als bei bisherigen Technologien von anderen Parteien zum Nachbau ausgenutzt werden könnte. Zudem erschwert die Überschneidung mit der zivilen KI-Forschung eine eindeutige Klärung der Absichten (Bostrom 2017).

Der Einsatz von KI in Waffensystemen könnte durch die Beschränkung der strategischen und taktischen Ziele des Systems und der damit verbundenen Handlungsoptionen geschehen (Kahl/Mölling 2005: 350). Solche Missionsziele und die Handlungen des Systems könnten in einer Art Flugschreiber aufgezeichnet werden. Eine von Gubrud und Altmann vorgeschlagene „Glass Box" könnte die Transparenz immens steigern:

> „A time slice of the data stream immediately prior to and including the selection and engagement commands could be designated as the primary record of the engagement. This record would be held by the state party, but a cryptographic code called a 'hash' of the record would be recorded by a 'glass box' […] together with a time stamp of the moment the engagement command was issued. The hash would serve as a digital seal of the engagement record; if even a single bit of the record were later altered, the hash would not match." (Gubrud/Altmann 2013: 6)

Sollte der Verdacht einer illegalen Kampfhandlung vorliegen, müsste der Staat die Aufzeichnungen der Box an ein internationales Verifikationsregime übergeben.

3.3 KONZEPTUELLE REFLEXION: STRATEGISCHE INSTABILITÄT UND PROLIFERATION

Nachdem die Relevanz und die Hürden von lernender KI als Kontrollgegenstand festgestellt wurden, wird in diesem Kapitel die Perspektive der Rüstungskontrolltheorie genutzt, um die Folgen des Einsatzes von KI abzuleiten. Das Ziel, zwischenstaatliche Beziehungen zu stabilisieren, soll erreicht werden, indem eine militärische Eskalation und ein Rüstungswettlauf verhindert werden. Doch die deeskalierenden Handlungsmöglichkeiten des Menschen, die Rüstungskontrolle zur Regulierung nutzt, würden mit einer Autonomisierung der Waffensysteme wegfallen. Zudem könnte KI zu einer vertikalen Proliferation – militär-technologische Weiterentwicklung und Verbesserung vorhandener Kapazitäten – und damit zu einem neuen Rüstungswettlauf beitragen. Auch können KI-Entwicklungen für zivile und militärische Verwendungszwecke in ähnlichem Maße genutzt werden und damit zu einer horizonta-

len Proliferation – Verbreitung von militär-technologischem Wissen und Waffensystemen innerhalb staatlicher und nichtstaatlicher Akteure – beitragen.

3.3.1 KRISENINSTABILITÄT: MANGEL DER DEESKALIERENDEN HANDLUNGSMÖGLICHKEITEN DES MENSCHEN

Wird eine KI mit hoch entwickelten Wahrnehmungs-, Lern- und Schlussfolgerungsfähigkeiten in Waffensystemen eingesetzt, so kann ein hoher Grad an Autonomie erreicht werden. Dieser gefährdet das allgemeine rüstungspolitische Ziel der Krisenstabilität, da der ausgleichende Faktor des Menschen minimiert wird. Dieser besteht darin, dass Menschen das Geschehen verlangsamen:

> „Despite modern communications and electronic data processing, officials are still limited by ordinary human intelligence, the conventional speed of spoken language, reading motions of the eye and the emotional accompaniments of responsibility in a crisis." (Schelling/Halperin 1961: 27)

In dieser Zeitspanne bieten sich dem Menschen drei deeskalierende Handlungsmöglichkeiten:

- Validierung der maschinellen Meldung oder Empfehlung
- Kommunikation mit dem Gegner für Verhandlungen oder Klärungen
- Abwägung der moralischen und rechtlichen Implikationen

Rüstungskontrolle setzt auf diese Handlungsmöglichkeiten, indem die Einsatzbereitschaft von Waffensystemen verlängert wird. "Many weapon limitations seem to be oriented, implicitly if not explicitly, towards the tempo of decision" (Schelling/Halperin 1961: 27). In der konventionellen Rüstungskontrolle geschieht dies dadurch, dass Raketen nicht dauerhaft abschussbereit sein dürfen, Sprengköpfe separat zu den Raketen gelagert werden oder mit Raketen bestückte U-Boote in Küstennähe bleiben müssen. Auch eine nukleare Eskalation wurde im Kalten Krieg in mehreren Fällen verhindert, indem Menschen einen technischen Fehlalarm identifizierten (Schlosser 2013). Die Rüstungskontrolle nutzt diesen menschlichen Faktor nicht nur, sondern stärkt ihn durch das Instrument der vertrauensbildenden Maßnahmen. Diese Maßnahmen zielen auf den Aufbau von Vertrauen zwischen Menschen unter anderem durch jeglichen Informationsaustausch, Zulassung fremder Beobachter bei Militärübungen, Austauschprogramme für Offiziere und Auszubildende oder direkte Kommunikationswege für Krisensituationen (Goldblat 2002: 11).

Ist die Autonomie eines Waffensystems auf einem Grad, der keine menschliche Überwachung und Intervention zulässt, so wird die automatische Eskalation einer Situation wahrscheinlicher. Beispielhaft für solch einen durch KI verursachten Zwischenfall steht der *Flash Crash* im Mai 2010 an der New Yorker Börse, als eine Marktmanipulation eine Abwärtsspirale von Verkäufen durch computergesteuerte Hochfrequenzhändler initiierte. Als Konsequenz etablierten die Behörden Sicherheitsmechanismen (CFTC/SEC 2010). Autonome Waffensysteme könnten in ähnliche Situationen gelangen, in denen ein Fehler oder ein Zufall eine außerordentlich schnelle Eskalation auslöst. Ein

Rückfallmechanismus auf einen menschlichen Entscheider bei nonkonformen Verhalten der KI wäre eine stabilisierende Vorkehrung, die eine gewaltsame Eskalation verhindern würde (Scharre 2016: 38–39). Solch ein Mechanismus ist auch notwendig, damit vertrauensbildende Maßnahmen nicht überflüssig werden: Das aufgebaute Vertrauen hat weniger Wert, wenn die Waffensysteme eigene Entscheidungen treffen und die menschlichen Einschätzungen der gegnerischen Partei keinen Einfluss mehr ausüben.

Doch insbesondere lernende KI bedarf menschlicher Überwachung, solange die systematischen Lernfehler und die Anfälligkeit für Manipulationen nicht behoben wurden (siehe Kapitel 2.2). Muss eine KI beispielweise zwischen Kombattanten und Zivilisten unterscheiden, so kann ein fehlerhaftes oder manipuliertes Lernmodell zu einer falschen Klassifikation führen. Feindliche Kombattanten könnten sich schützen, indem sie optische Merkmale an der Kleidung oder der Waffe so verändern, dass sie falsch klassifiziert werden. Lernende KI, dem Kernstück autonomer Waffensysteme, fehlt die deeskalierende Handlungsmöglichkeit des Menschen, die solch ein Szenario verhindern könnte. So soll hier nach Altmann und Sauer resümiert werden:

> „Speed is undoubtedly a tactical advantage on the battlefield, and humans are slower than machines. But strategic stability is essential for survival. When it comes under threat, some remainder of human slowness is a good thing." (Altmann/Sauer 2017: 136)

3.3.2 RÜSTUNGSWETTLAUFINSTABILITÄT UND VERTIKALE PROLIFERATION: RISIKO WIEDERBELEBTER AUSSENGELEITETER RÜSTUNGSDYNAMIK

Zwischen der Entwicklung von KI und der Rüstungsdynamik eines Landes kann eine Wechselwirkung entstehen, die sowohl von innen als auch von außen geleitet werden kann. Die Perspektive der innengeleiteten Rüstungsdynamik sieht den Ursprung der Wechselwirkung in innergesellschaftlichen Kräfteverhältnissen. Die KI-Forschung ist Teil einer größeren technischen militärischen Entwicklungsphase, die allen voran in den USA sichtbar ist (Neuneck/Alwardt 2008). Diese Aufrüstung wird vor allem in Demokratien vorangetrieben, um möglichst überlegene Waffensysteme zu entwickeln, die das Risiko eigener Opfer in Kriegshandlungen minimieren (Shaw 2005: 79; Schörnig 2008). Zusätzlich besitzen diese Staaten auch starke militärisch-industrielle Akteure, die ihren innenpolitischen Einfluss für eine stetige militärische Weiterentwicklung geltend machen (Müller/Schörnig 2006: 106).

Die Perspektive der außengeleiteten Rüstungsdynamik sieht den Ursprung der Rüstungsdynamik in der Beziehung zwischen zwei oder mehreren Staaten. In einer Form streben die Staaten nach einer intensiven militärtechnologischen Entwicklung, um eine technologische Überlegenheit gegenüber anderen Staaten zu erlangen. Matthew Evangelista (1988) zeigt anhand des nuklearen Rüstungswettstreits des Kalten Krieges, dass sich die Entwicklung technologischer Innovationen in Waffensystemen stark zwischen den USA und der UDSSR unterschieden. Während in den USA die Innovationen aus der starken Zivilgesellschaft heraus entsprangen, wurden Innovationen in der UDSSR zentralistisch und reaktiv vom Staat vorgegeben. Evangelistas Modell ist hilfreich zum Verständnis von aktuellen Innovationen in der KI-Forschung: Die US-amerikanische Forschung wird größtenteils

von zivilen Unternehmen finanziert und bestätigt damit den bottom-up-Ansatz (Bughin et al. 2017: 10). Russland und China hingegen forcieren die Entwicklung:

> „Artificial intelligence is the future, not only for Russia but for all humankind, […]. Whoever becomes the leader in this sphere will become the ruler of the world." – Vladimir Putin (Lant 2017)

Zudem verkündete das *Russian Military Industrial Committee*, dass bis zum Jahr 2025 30% der Militärtechnologie durch Robotik und autonome Systeme ersetzt werden sollen (Association of the United States Army 2017: 1). Auch China veröffentlichte Pläne, in denen es 150 Milliarden US Dollar in den kommenden Jahren verspricht, um China bis zum Jahr 2030 zum Innovationszentrum für KI zu machen:

> „We need to speed up building China into a strong country with advanced manufacturing, pushing for deep integration between the real economy and advanced technologies including internet, big data, and artificial intelligence." – Xi Jingping (Yu/Jing 2017).

Der außerordentliche Fortschritt der chinesischen KI-Forschung soll weiter ausgebaut werden und das Militär strebt zivile Kooperationen an, um KI in die Streitkräfte zu integrieren (Kania 2017). Die Investitionsprogramme von Russland und China könnten den militär-technologischen Vorsprung der USA verringern.

Sollten KI-gesteuerte Waffensysteme einen hohen strategischen Vorteil – wenn auch anderen Einsatz – vergleichbar der Atomwaffe erhalten und die angekündigten staatlichen Investitionen in militärische Anwendungen realisiert werden, könnte ein neuer Rüstungswettstreit zwischen mindestens den drei Akteuren China, Russland und USA initiiert werden.

3.3.3 DUAL-USE UND HORIZONTALE PROLIFERATION: UNKONTROLLIERBARE VERBREITUNG UND NUTZUNG

Ein weiteres Ziel der Rüstungskontrolle ist die Eindämmung der Verbreitung (Proliferation) von Waffen(-systemen) und militär-technischem Wissen. Dieses Ziel wird vor allem durch das Problem doppelter Nutzung – *Dual-Use*[6] – gefährdet, da die fortschreitende zivile Entwicklung für militärische Zwecke genutzt werden kann:

6 „The trade in dual-use items – goods, software and technology that can be used for both civilian and military applications and/or can contribute to the proliferation of Weapons of Mass Destruction (WMD) – is subject to controls to prevent the risks that these items may pose for international security." (European Commission 2017).

> „[O]pen-sourcing the code for autonomous weapons seems undesirable, and we have not heard anybody calling for that to be done. But basic research in AI is typically not application-specific in this way. Rather, to the extent that it succeeds, it will deliver algorithms and techniques that could be used in a very wide range of applications." (Bostrom 2017: 137)

Selbst wenn die Quellcodes zukünftiger (semi-)autonomer Waffensysteme nicht öffentlich zugänglich sein werden, so ist doch die öffentliche Grundlagenforschung[7] nicht anwendungsgebunden und könnte für illegale Zwecke instrumentalisiert werden. Eine Vielzahl an KI-Anwendungen und Programmiergerüsten sind frei verfügbar.[8] Diese Open-Source-Projekte werden einerseits zur Entwicklung in der Rüstungsindustrie und andererseits in der Entwicklung ziviler Projekte genutzt. Je vielfältiger die Anwendungsgebiete einer einzigen zivil entwickelten KI sind, desto wahrscheinlicher ist ihre Nutzung in Waffensystemen. Im Falle der Nuklearwaffentechnologie konnte die Verbreitung des Wissens durch die Teststoppnorm beschränkt werden, da die Entwicklung einer fortgeschrittenen nuklearen Waffe mehrere Testläufe benötigt. Die Tests müssen in einem Maßstab durchgeführt werden, welcher unweigerlich aufspürbare seismische Wellen, hydroakustische Signale oder Radionuklide emittiert. Für KI wäre ein Verbot von Tests nicht zielführend, da Funktionstests in kleinem Maßstab oder in Simulationen – welche keine aufspürbaren Signale aussenden – durchgeführt und erst anschließend auf viele Systeme skaliert werden können. Die Weiterentwicklung grundlegender KI zu einer waffenfähigen Anwendung ist eine vergleichsweise kleine Hürde und lässt das Ziel von Rüstungskontrolle, die Verbreitung risikoreicher Technologie zu verhindern, in weite Ferne rücken.

4. LERNENDE KÜNSTLICHE INTELLIGENZ IN VERIFIKATIONSMASSNAHMEN

„Trust, but verify" ist ein oftmals zitiertes Diktum in der Rüstungskontrolle. Denn eine vereinbarte Beschränkung von Rüstung befreit einen Staat noch nicht vom grundsätzlichen Misstrauen gegenüber anderen Staaten. Erst die Verifikation – also die Überprüfung, ob Staaten einen Rüstungskontrollvertrag einhalten – vermag das Misstrauen zu mindern und stärkt das Streben eines jeden Staates nach Sicherheit.

Die drei Ziele der Verifikation sind, *erstens*, Transparenz zu schaffen und damit Vertragsbrüche frühzeitig zu erkennen, um diplomatische, militärische oder ökonomische Maßnahmen einzuleiten. Durch die Aussicht auf diese Reaktionen sollen Verifikationsmaßnahmen auch vor einem Bruch abschrecken. Neben der abschreckenden Funktion sollen diese Maßnahmen, *zweitens*, auch Vertrauen zwischen den Parteien aufbauen. Die Bestätigung, dass alle Mitgliedsstaaten den Vertrag einhalten, schafft Vertrauen in den Nutzen der Rüstungskontrolle für den Schutz der nationalen Interessen. *Drittens* ermöglichen Verifikationsmaßnahmen fälschlich beschuldigten Staaten, ihre Vertragskonformität zu demonstrieren (Goldblat 2002: 309). Wenn solch eine Anschuldigung formuliert wird, müssen die Beweismittel aus Verifikationsmaßnahmen und geheimdienstlichen Quellen gesammelt und be-

7 Auf https://arXiv.org veröffentlicht ein Großteil namhafter Forscher ihre neusten Ergebnisse (ohne Quellcode).

8 Auswahl frei zugänglicher Lernarchitekturen: Apache Singa, H2O, TensorFlow, Torch, Accord.NET.

wertet werden. Sollte sich die Anschuldigung bewahrheiten, können die Mitgliedsstaaten Maßnahmen innerhalb des Regimes beschließen oder den Vertragsbruch an den Sicherheitsrat der Vereinten Nationen weiterleiten (Müller/Schörnig 2006: 150–153). Wenn diplomatische Maßnahmen zur Richtigstellung des Vertragsbruchs fehlschlagen, können Embargos, Sanktionen oder die Androhung militärischer Gewalt folgen.

Durch die potentiellen Maßnahmen wird deutlich, dass die Validität und Qualität der gesammelten Beweismittel essentiell für die Beurteilung sind. Durch Überwachungen aus der Ferne (Satelliten, Flugzeuge, Radar und andere Sensorsysteme) und Inspektionen vor Ort werden Streitkräfte, Waffen oder Aktivitäten der Mitgliedsstaaten überprüft. Die Inspektionen können kontinuierlich, periodisch oder ad-hoc geschehen. Der Verifikationsprozess beinhaltet die Datensammlung, -verarbeitung und -analyse, um verwertbare Informationen ableiten zu können (Goldblat 2002: 310). In einem oder allen dreien dieser Schritte kann lernende KI eingesetzt werden, um die Validität und Qualität zu erhöhen oder die Effizienz von menschlichen Analysten zu steigern.

4.1 ANWENDUNGSGEBIETE UND TECHNISCHES POTENTIAL

Das Potential von lernender KI zur Datensammlung, -verarbeitung und -analyse für Verifikationsmaßnahmen lässt sich bereits heute in Anwendungen oder Forschungen erkennen. Für eine Betrachtung dieser Fälle eignen sich verschiedene Datenquellen als Ausgangspunkte: Satellitenbilder aus dem Weltraum, Inspektionen am Boden oder Sensoren in globalen Netzen.

4.1.1 AUS DEM WELTRAUM: MASCHINELLE FERNERKUNDUNG MIT SATELLITENAUFNAHMEN

Die Analyse von Satellitenaufnahmen spielt beispielweise in den Rüstungskontrollregimen zur Nichtverbreitung von Kernwaffen und -energie oder Beschränkung von Kurz- und Mittelstreckenraketen eine wichtige Rolle. Die aus Satellitenbildern gewonnenen Informationen – der Prozess ist auch bekannt als Fernerkundung – werden genutzt, um Vertragsbestimmungen zu verifizieren (Patton et al. 2016; Niemeyer/Ruthowski 2016). Um die Validität und Aussagekraft dieser Analysen zu erhöhen, kann lernende KI zur Identifikation von Objekten und zeitlichen Veränderungen auf den Aufnahmen eingesetzt werden.

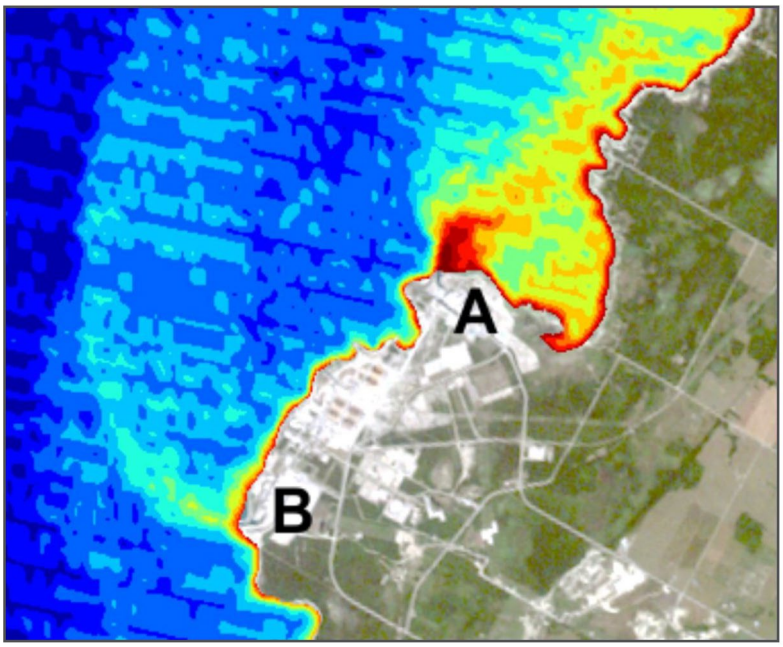


Abb. 2: Eine Thermografie des Kernkraftwerks Bruce, Kanada zeigt, dass Block A im Gegensatz zu Block B in Betrieb ist. (Truong et al. 2005: 3)

Eine intensive Anwendung von Luftaufnahmen findet sich beispielsweise bei der *Internationalen Atomenergie-Organisation* (IAEO), die die Einhaltung des *Nuklearen Nichtverbreitungsvertrages* kontrolliert. Für die Überwachung nuklearer Einrichtungen nutzt die IAEO die räumliche, zeitliche und multispektrale Dimension, die ihr die Satellitenaufnahmen bieten. Das Ziel der Analysen ist es, nicht deklarierte Produktionsanlagen von hochangereichertem Uran oder Plutonium zu detektieren, die Nutzung und Verarbeitung der Schwermetalle nachzuvollziehen und Objekte von Interesse für Vor-Ort-Inspektionen zu identifizieren. Dafür werden Indizien wie beispielsweise die Temperaturabstrahlung oder die optischen Veränderungen der Gebäude genutzt, um den Ausbau oder die Nutzungsintensität einer Anlage zu überwachen (Truong et al. 2005; Johnson et al. 2014).

Eine konkrete Herausforderung liegt beispielsweise in der Identifikation und Beobachtung von Uranmühlen, die im Kernbrennstoffkreislauf aus natürlichem Uran den so genannten *yellow cake* herstellen, ein Vorprodukt des in Atomkraftwerken oder Nuklearwaffen genutzten Urans. Da diese auf Satellitenaufnahmen sehr ähnlich wie völlig harmlose Kupfermühlen aussehen, nutzt eine von Forschern entwickelte KI Charakteristika einzelner Gebäude und die Größe des Komplexes, um die Anlagen zu klassifizieren (Sundaresan et al. 2017). Die Klassifizierung führt die KI mithilfe eines durch Menschen vordefinierten Entscheidungsbaums durch. Obwohl die IAEO ihre Maßnahmen in der Vergangenheit immer an die technologische Entwicklung angepasst hat, gibt es derzeit keine Hinweise,

dass sie lernende KI in der praktischen Analyse von Luftaufnahmen nutzt.[9] Sie wird hingegen weiter als nicht erschlossene Innovation in der aktualisierten Forschungsplanung von 2018 gelistet – jedoch ohne hohe Priorität (International Atomic Energy Agency 2018: 15).

Kommerzielle Anbieter offerieren bereits nutzbare Anwendungen: Das Geoinformationssystem ENVI vertreibt ein Modul, das Forschern und Analysten ermöglicht, ein Beispielbild eines gesuchten Objekts, z.B. Tanks für chemische Substanzen, als Trainingsdaten einzuspeisen und das Programm anschließend danach auf Luftaufnahmen suchen zu lassen. Neben Anwendungsgebieten in Stadtplanung, Naturschutz und Forstwirtschaft wirbt der Hersteller explizit mit der Auffindung von militärischen Fahrzeugen, Landezonen oder Gebäuden (Harris Cooperation 2017). Auch andere Anbieter[10] verfolgen das Geschäftsmodell der KI-basierten Auswertung kommerzieller Satellitenbilder. „*We could not have done this five years ago*", berichtet der CEO Pavel Machalek von SpaceKnow (Dillow 2016). Und das Potenzial sei weiterhin groß, denn die Fortschritte in der Kombination aus Rechenleistung, maschinellem Lernen und Satellitenaufnahmen seien gerade erst am Anfang.

Zwar stehen nichtstaatlichen Akteuren aufgrund amerikanischer Restriktionen nur relativ grob aufgelöste Satellitenbilder mit einer Auflösung von 30 bis 40 cm pro Pixel zu Verfügung (Shalal 2014), während US-Spionagesatelliten eine vermutete Auflösung von 15 cm haben (Krebs 2017). Dennoch ergeben sich enorme Chancen zur Nutzung dieser Aufnahmen durch lernende KI, da die KI den Nachteil schlechterer Auflösung zu bedeutenden Teilen ausgleichen kann. Abgesehen von Klein- und Leichtwaffen kann kommerzielle lernende KI militärisches Großgerät auch auf den grobauflösenden kommerziellen Satellitenaufnahmen identifizieren.

Je nach Größe des beobachteten Objekts kann eine lernende KI auch eindeutige Merkmale erkennen und das Objekt über mehrere Aufnahmen hinweg verfolgen. Auffällige Veränderungen an nuklearen, chemischen oder biologischen Fabriken können automatisch analysiert werden und Anomalien den Analysten der Kontrollregime mitgeteilt werden. Durch diese herausragende Anomalie- und Mustererkennung kann der Einsatz von lernender KI der satellitengestützten Verifikation zu einer erhöhten Validität verhelfen und sie somit als Instrument stärken.

4.1.2 AM BODEN: WAFFENHANDEL UND ZIVILE RÄUMUNG NICHTDETONIERTER SPRENGKÖRPER

Auch der Handel mit konventionellen Waffen und damit u.a. die Einhaltung des *Vertrags über den Waffenhandel* könnte mithilfe von lernender KI kontrolliert werden. Der Vertrag verpflichtet Staaten, Waffenexporte bis an seine Ziele zu verfolgen und einen möglichen Transfer an Staaten, welche Menschrechtsverletzungen oder Verletzungen des humanitären Völkerrechts begehen, zu verhindern. Um

9 Innerhalb des neu entwickelten Geospatial Exploitation System der IAEO sollen auch Analysetools kommerzieller Anbieter zu Verfügung stehen (Balter 2014: 6). Dort könnte durchaus maschinelles Lernen zum Einsatz kommen.

10 Der Satellitenbetreiber DigitalGlobe bietet ein Programmiergerüst für Objektauffindung an: http://deepcore.io/. Die Startups SpaceKnow (https://spaceknow.com/), Orbital Insight (https://orbitalinsight.com/) und Descartes Labs (https://www.descarteslabs.com/) bieten Online-Plattformen zur Objekt- und Mustererkennung an. Auf der Unterseite https://search.descarteslabs.com/ kann eine vereinfachte Funktion getestet werden.

dies auch für Transitgüter gewährleisten zu können, müssen die unzähligen Frachtcontainer kontrolliert werden, die täglich in Häfen ein-, aus- und umgeladen werden (Holtom/Bromley 2011). Aufgrund des immensen Handelsvolumens kann eine menschliche Kontrolle bestenfalls stichprobenartig erfolgen und basiert auf Risikofaktoren wie Herkunft, Destination und angegebenem Inhalt. Doch ein KI-System wäre in der Lage, mithilfe von Röntgenaufnahmen Rüstungsgüter innerhalb unzähliger verschlossener Container zu identifizieren, wie es ein Forscherteam des University College London bereits demonstrierte (Jaccard et al. 2016). Zwar wären Menschen dazu in der Lage, die durch die Sensoren kreierten Abbilder des Inhalts zu interpretieren, doch kann maschinelles Lernen die Daten wesentlich schneller erfassen, Muster in den Daten erkennen und eine Wahrscheinlichkeit für die Art des Inhalts ausgeben. Die KI müsste in der Trainingsphase des Lernmodells zahlreichen verschiedenen Szenarien gegenübergestellt werden und zusätzlich mit der richtigen Antwort – bspw. „Sprengkopf“ oder „kein Sprengkopf“ – versorgt werden.

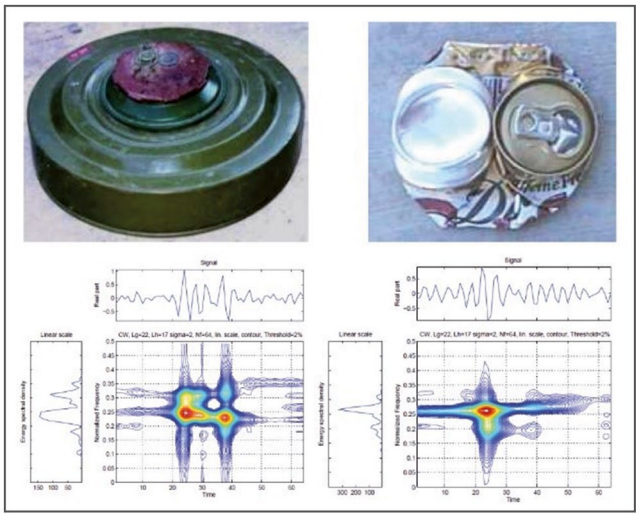


Abb. 3: Vergleich von Sensordaten zwischen Antipersonenmine und Getränkedose (Sun/Li 2005: 3)

Ein zweites Anwendungsbeispiel stammt aus der humanitären Rüstungskontrolle: Antipersonenminen und Streumunition verbleiben auch nach Konfliktende im Boden und gefährden die Zivilbevölkerung. Die humanitären Rüstungskontrollverträge, *Ottawa-Konvention* und *Übereinkommen über Streumunition*, schreiben neben einem allgemeinen Verbot auch die Räumung von bereits platzierten Waffen durch die Konfliktparteien vor. Dieser aufwendige Prozess verlangt ausgebildete Teams, die systematisch die Gebiete abstecken, mit Metalldetektoren nicht detonierte Minen oder Munition aufspüren und anschließend kontrolliert sprengen. Für die große Herausforderung, die Objekte aufzuspüren, wurden und werden möglichst effiziente Methoden entwickelt. Ein vielversprechender neuer Ansatz kombiniert einen Bodenradar mit maschinellem Lernen. Wie in der Abbildung zu sehen ist, lassen sich Radarbilder einer Antipersonenmine und einer platten Getränkedose rein visuell nicht leicht unterscheiden. Zu diesem Zweck trainierten verschiedene Entwicklerteams eine lernende KI auf solche Eingabedaten (Núñez-Nieto et al. 2014; Seiffert et al. 2013). Diese Methode ermöglicht es, das System auf einen Geländewagen zu befestigen und einige Quadratmeter vor dem Wagen den Bo-

den zu scannen. Dabei ist das Ziel, die Quote von Unfällen oder Fehlalarmen im Vergleich zu anderen Aufspürmethoden zu verringern.

Die zwei Fälle dieses Kapitels haben gemeinsam, dass die KI unmittelbare Sensordaten aus der Umwelt erhält, um Muster zu erkennen. Bei der Aufspürung von Rüstungsgütern oder Antipersonenminen liegt das Interesse in der Unterscheidung zu anderen Objekten auf Basis von unvollständigen oder unstrukturierten Daten. Da an diesen Fällen bereits aktiv geforscht wird, ist eine Realisierung und Anwendung wahrscheinlich, wenn KI-Systeme eine bessere Effizienz gegenüber alten Systemen beweisen können. Werden diese Entwicklungen verallgemeinert und extrapoliert, so zeigt sich ein erhebliches Potential für materielle Transparenz: Aktivitäten oder Objekte, die gegen Rüstungskontrollabkommen verstoßen, können auch bei größter Sorgfalt nur unvollständig gegen Entdeckung abgeschirmt werden. Die wenigen Informationen, die ungewollt durch die Abschirmung dringen, können entweder gar nicht oder zu langsam von Menschen oder statistischen Modellen interpretiert werden, doch lernende KI kann an dieser Stelle neue Transparenz schaffen.

4.1.3 DURCH VIELFÄLTIGE SENSOREN: „MEASUREMENT AND SIGNATURE INTELLIGENCE"

Die Verarbeitung von elektrooptischen, elektromagnetischen, akustischen und geophysikalischen Daten sowie nuklearen, biologischen und chemischen Spurenelementen wird in der nachrichtendienstlichen Arbeit „measurement and signature intelligence" (MASINT) genannt (Aid 2014: 120–122). Die grundlegende Idee von MASINT ist, dass das Aufklärungsobjekt über verschiedene Informationsträger (Strahlung, Schall, Stoffe usw.) hinweg eine eindeutige Signatur absondert. Die Analyse dieser Informationen kann für Verifikationsmaßnahmen genutzt werden.

Eine Kontrollmethode, die akustische und seismische Sensordaten nutzt, wurde durch das Bochumer Verifikationsprojekt entwickelt. Die im Projekt entwickelten Messgeräte wurden in Abständen von bis zu 200m platziert, um eine Linie an Sensoren zu haben, die auch Fahrzeuge abseits der Straßen aufspüren. Die Messgeräte konnten vorbeifahrende Militärfahrzeuge anhand von akustischen und seismischen Signalen in verschiedene Fahrzeugkategorien einordnen (Hochmuth 2003). Die erfolgversprechendere akustische Erkennung konnte zwischen jeweils fünf Ketten- und Räderfahrzeugen anhand des Motorschalls unterscheiden (Altmann et al. 2002). Die variierende Erfolgsrate von 69% bis 98% könnte mit maschinellem Lernen auf einem hohen Niveau stabilisiert werden und die Resilienz gegen Störgeräusche, wie beispielsweise einem Regenschauer, verbessert werden. Eine Linie solcher Messgeräte könnte beispielsweise die quantitative Limitierung eines bestimmten Fahrzeugtyps kontrollieren. Die Identifizierung von individuellen Fahrzeugen dürfte dagegen schwer sein. Zwar variiert der Motorschall je nach Fahrzeug in geringem Maße, allerdings kann dieser sich durch Abnutzung oder Reparaturen verändern und damit wäre eine Wiedererkennung nicht möglich.

In einem weiteren vielversprechenden Anwendungsgebiet soll lernende KI zur Verifikation der Einhaltung des *Kernwaffenteststopp-Vertrags* (CTBT) genutzt werden. Das bereits bestehende internationale Überwachungssystem (*International Monitoring System* – IMS) steht beispielhaft für eine herkömmliche Verarbeitung von massenhaften Sensordaten. Durch die Erfassung dieser in einem

weltweiten Netz aus Messstationen kann das IMS nukleare Explosionen diagnostizieren. Seismographen, Hydroakustiksensoren, Infraschallstationen und Radionukliddetektoren generieren große Mengen an Daten. Die zentrale Datenverarbeitung in Wien analysiert und reduziert die rohen Sensordaten, um aus dem ständigen Rauschen Signale zu erkennen und in „nukleare Explosion" und „keine nukleare Explosion" zu klassifizieren (Russell et al. 2010: 32). Der Analysevorgang kann zurzeit noch nicht völlig automatisch durchgeführt werden. Die Analysten müssen das Ergebnis der automatischen Systeme noch aufwändig nachbearbeiten, da die automatische Bearbeitung durch starkes Rauschen, inkorrekte Klassifikation oder falsche Assoziationen fehleranfällig ist. Um die Ergebnisse weiter zu optimieren, wurden schon 2009 auf der *International Scientific Studies* (ISS) Konferenz in Wien von verschiedenen Projektgruppen neue Methoden vorgeschlagen. Die Projektgruppen waren sich in der Problemlage einig: Die durch die Sensoren des IMS aufgezeichneten Daten sind zu komplex, um eine gängige statistische Verarbeitung durch eine lineare Diskriminierung – eine lineare Funktion, die die Grenze zwischen Gruppen definiert – durchzuführen. Für eine nicht lineare Diskriminierung testen alle vier Projektgruppen die Fähigkeit des maschinellen Lernens. Die Projektgruppen nutzten bereits klassifizierte Daten aus den letzten zehn Jahren als „Ground Truth" (Trainingsset für den maschinellen Lernalgorithmus). Die Prototypen analysierten seismische Daten (Kleiner et al. 2009), hydroakustische Daten (Tuma/Igel 2009) und Infraschalldaten (Procopio et al. 2009) sowie Messdaten von Radionukliden (Stocki et al. 2010). Alle Prototypen waren in der Lage, Erleichterungen für die Analysten und akkuratere automatisierte Auswertungen vorzuweisen. Dennoch kam man damals zu dem Schluss, die Systeme seien noch nicht einsatzbereit und müssten verfeinert werden. Insgesamt haben diese Forschungsprojekte und eine verbesserte Weiterentwicklung (Arora et al. 2013) gezeigt, welch hohes Potential in lernender KI für die Verifikation in der Rüstungskontrolle in den IMS-Daten steckt.

Die Analyse großer Datenmengen aus der physischen Welt – übersetzt durch Sensoren – kann von lernender KI im Vergleich zu menschlichen Analysten in hohem Maße verbessert werden. Speziell die Kontrolle des nuklearen Teststopps wird auf kurze Sicht enorm von maschinellem Lernen profitieren können, da die Daten durch die Sensoren bereits strukturiert und damit leichter analysierbar sind. Lernende KI macht diese Art der Verifikation erheblich attraktiver und könnte dazu führen, dass andere Waffen auf eine ähnliche Weise kontrolliert werden.

4.2 KONZEPTUELLE REFLEXION: VERBESSERUNG BESTEHENDER METHODEN

Die Fälle des vorherigen Kapitels zeigen, dass lernende KI die Fähigkeiten besitzt, um Rüstungskontrolle, insbesondere deren Verifikationsmaßnahmen, zu stärken. Seit Ende des Ost-West-Konflikts stagniert die Bedeutung von Verifikationsmaßnahmen aufgrund von neuartigen Waffentechnologien und vereinzelten Vertragsbrüchen (Pilat 2002: 85–87). Lernende KI hätte das Potential, diesen Maßnahmen wieder zu erhöhter Relevanz zu verhelfen, sofern die Staaten den entsprechenden politischen Willen aufbringen würden, hier Anstrengungen zu unternehmen.

Die technische Weiterentwicklung der Verifikationsmaßnahmen würde, sofern umgesetzt, eine höhere Transparenz für die Überwachung der Kontrollgegenstände ermöglichen. Setzt sich die Er-

kenntnis einer verbesserten Verifikation durch, so wird auch das Ziel der Abschreckung vor einem heimlichen Vertragsbruch gestärkt. Die zwei weiteren Ziele von Verifikationsmaßnahmen – Vertrauensbildung und Demonstration von Vertragskonformität – können ebenfalls von der erhöhten Transparenz durch den Einsatz von lernender KI profitieren.

Wie in den vorangegangenen Kapiteln gezeigt wurde, ist der Einsatz von lernender KI in Machbarkeitsstudien bereits erprobt und in Verifikationsmaßnahmen wirksam. Gleichwohl stehen die lernenden KI-Programme aber auch noch vor großen Herausforderungen. Sie müssen gegebenenfalls – je nach Form der Eingabedaten – visuelle und andere sensorische Eindrücke „wahrnehmen" und interpretieren. In den digitalen Daten müssen sie Muster erkennen und mithilfe des vorhandenen Wissens ein Analyseergebnis schlussfolgern. Wahrnehmung, Mustererkennung und Generierung von Wissen können durch maschinelles Lernen verbessert werden, da menschliche Vorgaben an das Analyseprogramm die notwendige Komplexität nicht erreichen könnten. Die Anpassungsfähigkeit maschinellen Lernens ermöglicht den Programmen eine Entwicklung parallel zu den Kontrollgegenständen, sollten sich Charakteristika der analysierten Objekte über die Zeit verändern. Lernende KI ist in der Lage, aus der hoch komplexen und wandelbaren Umwelt der Verifikationsmaßnahmen Erkenntnisse zu destillieren, welche die rüstungspolitische Transparenz erhöhen. Zusätzlich erlaubt lernende KI eine höhere Skalierbarkeit – eine deutliche Erhöhung der Zahl der durchgeführten Analysen:

> „If we were to attempt to manually exploit the commercial satellite imagery we expect to have over the next 20 years, we would need eight million imagery analysts." – Robert Cardillo, Direktor der National Geospatial-Intelligence Agency (Cardillo 2017)

Zugleich bietet lernende KI weitere Vorteile gegenüber menschlichen Analysten:

> „Technology does have advantages over human inspectors. It can operate continuously and at a constant level of observation. Its data is readily comparable. It can be limited to detecting treaty-relevant information, while ignoring other types of information." (UNIDIR/VERTIC 2003: 27)

Doch ob lernende KI als offizielles Verifikationsinstrument in der traditionellen Rüstungskontrolle genutzt wird, ist eine politische Entscheidung der Mitgliedsstaaten. Aufgrund von politischen Erwägungen wurden bereits technische Verifikationsmöglichkeiten künstlich begrenzt:

> „The INF Treaty […] permitted an x-ray to be taken of missile canisters to determine the type of missile inside, but the machinery was set to a certain resolution so that sensitive design information could not be obtained." (UNIDIR/VERTIC 2003: 27)

Auch die technischen Maßnahmen des *Open-Skies-Vertrags* – ein Vertrag, der nicht auf Waffenbegrenzung, sondern Vertrauensbildung zielt – wurden absichtlich auf einen niedrigeren Standard

begrenzt: Während der vertraglich zugesicherten Überflüge dürfen Luftbildaufnahmen mit einer maximalen Auflösung von 30 Zentimetern vorgenommen werden, obwohl kommerzielle Satellitenbilder bereits die gleiche oder eine besser Auflösung bereitstellen. Auch bei der Modernisierung von analogen auf digitale Aufnahmegeräte wurde darauf geachtet, die maximal vertraglich vereinbarte Auflösung nicht zu überschreiten (Britting/Spitzer 2005). Technische Weiterentwicklungen an traditionellen Rüstungskontrollverträgen sind nicht im Interesse jener Nationen, die bereits einen Vorsprung in der Informationsgewinnung besitzen. Der Einsatz lernender KI kann – wie bei anderen traditionellen Verifikationsmaßnahmen bereits geschehen – durch politischen Willen oder militärische Geheimhaltungsinteressen verhindert oder begrenzt werden.

In der humanitären Rüstungskontrolle sind entwicklungshemmende Interessen von Nationalstaaten eher nicht zu erwarten. Pilotprojekte zur Aufspürung von Antipersonenminen und Streumunition (s. Kapitel 4.1.2) zeigen den immensen Nutzen für eine schnelle Räumung solcher nichtdiskriminierenden Waffen. Dieses Prinzip könnte auch von Akteuren eingesetzt werden, die den internationalen Handel von Kleinwaffen überwachen. Beispielsweise verfolgt das Projekt *itrace* durch Ermittlungen vor Ort den Verbreitungsweg von solchen Waffen. Die Ermittler könnten Unterstützung durch lernende KI erhalten, indem die KI Handelsmuster im Datenmaterial oder eindeutige visuelle Merkmale identifiziert und damit Einzelwaffen zu bestimmten Waffenbeständen zuordnen kann.

5. LERNENDE KI ZUGLEICH TEIL DES PROBLEMS UND DER LÖSUNG

In diesem Report wurde deutlich, dass in der Kontrolle von konventionellen Waffen, Massenvernichtungswaffen und modernen Waffensystemen lernende KI sowohl Teil des Problems als auch Teil der Lösung sein kann. Lernende KI als Kontrollgegenstand entzieht sich ihrem Wesen nach vielen Ansatzpunkten zur qualitativen oder quantitativen Beschränkung. Damit erhöht sie die Relevanz von alternativen Methoden, die auf eine gesamtmilitärische Transparenz oder auf vertrauensbildende Maßnahmen zielen. Genau in diesen Methoden kann lernende KI wiederum als verifizierendes Instrument eingesetzt werden. Durch eine präzise und umfangreiche Informationsverarbeitung kann lernende KI erhöhte Transparenz schaffen, die Einhaltung von Verträgen verifizieren und damit Vertrauen zwischen den Parteien stärken.

Aus der Perspektive der Rüstungskontrolltheorie hat der Einsatz lernender KI in Waffensystemen und Verifikationsmaßnahmen eine fragile strategische Stabilität zur Folge. Sollte lernende KI in Waffensystemen vermehrt eingesetzt werden, so kann sie die strategische Stabilität gefährden, indem der deeskalierende Charakter des Menschen minimiert, ein technologisches Wettrüsten gefördert und die Technologie unkontrolliert verbreitet wird. Sowohl diese theoretischen Überlegungen als auch die Erkenntnis, dass KI das Kernelement zukünftiger autonomer Waffensysteme sein wird, drängen zu einer Limitierung der Technologie durch die Rüstungskontrolle. Doch eine qualitative oder quantitative Kontrolle aufgrund des materiellen Äußeren, der äußerlich sichtbaren Fähigkeiten oder der inneren Funktionsweise ist im Falle (lernender) KI nicht möglich. Die traditionellen Ansatzpunkte sind an dieser Stelle erschöpft und es bleiben die Möglichkeiten, die KI während der Entwicklung oder des Einsatzes zu überwachen und zu limitieren. Sollte die Kontrolle des Einsatzes über eine „Glass

Box“ oder die präventive Kontrolle des Entwicklungsprozesses nicht durchsetzbar sein, könnten auch qualitative Lösungen auf einer höheren Ebene angestrebt werden. Die gesamtmilitärischen Fähigkeiten und Strategien könnten transparenter dargelegt und vertrauensbildende Maßnahmen intensiviert werden (Schörnig 2015).

Wird lernende KI in Verifikationsmaßnahmen eingesetzt, so kann sie die strategische Stabilität verbessern, da zu erwarten ist, dass sie die technischen Mittel zur Überwachung zu einer weitaus präziseren und umfangreicheren Informationsverarbeitung befähigt. Diese Chance bietet sich allerdings nicht ohne Hindernisse: Im gegenwärtigen Entwicklungsstand von lernender KI müssen die Transparenz und der Schutz gegen externe Manipulation verbessert werden, um als valide Verifikationsmethode zu gelten. Diese technischen Voraussetzungen müssen gegeben sein, damit das Vertrauen in die Methode und damit auch zwischen den Staaten aufgebaut werden kann. Das Potential von lernender KI als Verifikationsmethode lässt sich an zahlreichen Prototypen und ersten Anwendungen erkennen. Diese zeigen, dass die Fähigkeit des maschinellen Lernens in der Analyse von optischen, thermischen und topographischen Satellitenaufnahmen, Sensordaten in Vor-Ort-Inspektionen oder zivilen Waffenräumgen und unstrukturierten Datenmengen aus einem Netz an Messstationen eingesetzt werden kann.

Wann die zwei untersuchten Phänomene signifikante Auswirkungen auf die Rüstungskontrolle haben werden, ist von der Umwelt, in der die KI agieren soll, abhängig. Eine Typologisierung der Umwelt (Russell/Norvig 2010: 46) lässt erkennen, dass Handlungen in der realen Umwelt wesentlich mehr nicht vorhersagbare Faktoren der Umgebung einbeziehen müssen, als in einer digitalen Umwelt. Um Waffensystemen autonome Handlungen in der realen Umwelt zu ermöglichen, benötigt die KI-Forschung noch weitere Jahre. Doch kann lernende KI bereits für Verifikationsmaßnahmen genutzt werden, da die zu analysierenden Daten bereits in digitaler Form vorhanden sind oder nach einem vorgegebenen Schema interpretiert werden können, ohne auf plötzliche Änderungen in der Umwelt vorbereitet sein zu müssen. Die zeitige Einsatzfähigkeit ist von großer Bedeutung, da lernende KI bereits jetzt der Rüstungskontrolle zu neuen Kapazitäten verhelfen kann, bevor diese der neuen Herausforderung KI-gesteuerter Waffensysteme gegenübersteht.

Adams, Eric 2016: Why Only Isreal Can Customize America's F-35 (at Least for Now), https://www.wired.com/2016/05/israel-can-customize-americas-f-35-least-now/; 04.12.2017.

Aid, Matthew M. 2014: Measurement and Signature Intelligence, in: Dover, Robert/Goodman, Michael S./Hillebrand, Claudia (Hg.): Routledge Companion to Intelligence Studies, London, 114–122.

Algorithm Watch 2017: Antworten auf den Fragenkatalog für das Fachgespräch zum Thema „Künstliche Intelligenz" des Ausschusses Digitale Agenda am 22. März 2017.

Altmann, Jürgen 2008: Präventive Rüstungskontrolle, in: Becker, Una/Müller, Harald (Hg.): Rüstungskontrolle im 21. Jahrhundert, Berlin, 105–125.

Altmann, Jürgen/Linev, Sergey/Weiß, Axel 2002: Acoustic–seismic Detection and Classification of Military Vehicles – Developing Tools for Disarmament and Peace-keeping, in: Applied Acoustics 63: 10, 1085–1107.

Altmann, Jürgen/Sauer, Frank 2017: Autonomous Weapon Systems and Strategic Stability, in: Survival 59: 5, 117–142.

Arbatov, Alexei 2015: An Unnoticed Crisis. The End of History for Nuclear Arms Control?, https://carnegie.ru/2015/06/16/unnoticed-crisis-end-of-history-for-nuclear-arms-control-pub-60408; 19.07.2019.

Arora, N. S./Russell, S./Sudderth, E. 2013: NET-VISA. Network Processing Vertically Integrated Seismic Analysis, in: Bulletin of the Seismological Society of America 103: 2A, 709–729.

Artificial General Intelligence Sentinel Initiative 2017: A Working List. Definitions of Artificial Intelligence and Human Intelligence.

Association of the United States Army 2017: Integrating Army Robotics and Autonomous Systems to Fight and Win, https://www.ausa.org/publications/integrating-army-robotics-and-autonomous-systems; 19.07.2019.

Balter, E. 2014: Digital Declarations: The Provision of Site Maps under INFCIRC/540 Article 2.a. (iii), https://conferences.iaea.org/indico/event/47/contributions/8862/contribution.pdf.

Ben-Ari, Mordechai/Mondada, Francesco 2018: Elements of Robotics, Cham.

Bennaceur, Amel/Issarny, Valérie/Sykes, Daniel/Howar, Falk/Isberner, Malte/Steffen, Bernhard/Johansson, Richard/Moschitti, Alessandro 2013: Machine Learning for Emergent Middleware, in: Moschitti, Alessandro/Plank, Barbara (Hg.): Trustworthy Eternal Systems via Evolving Software, Data and Knowledge. Second International Workshop, EternalS 2012, Montpellier, France, August 28, 2012, Revised Selected Papers, Berlin, Heidelberg, 16–29.

Bojarski, Mariusz/Del Testa, Davide/Dworakowski, Daniel/Firner, Bernhard/Flepp, Beat/Goyal, Prasoon/Jackel, Lawrence D./Monfort, Mathew/Muller, Urs/Zhang, Jiakai/Zhang, Xin/Zhao, Jake/Zieba, Karol 2016: End to End Learning for Self-Driving Cars, https://arxiv.org/pdf/1604.07316v1.pdf; 19.07.2019.

Boston Dynamics 2018: Boston Dynamics. Changing Your Idea of What Robots Can Do, https://www.bostondynamics.com/robots; 05.01.2018.

Bostrom, Nick 2017: Strategic Implications of Openness in AI Development, in: Global Policy 8: 2, 135–148.

Boulanin, Vincent/Verbruggen, Maaike 2017: Mapping the Development of Autonomy in Weapon Systems, Stockholm, https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf; 19.07.2019

Britting, Ernst/Spitzer, Hartwig 2005: Der Open-Skies-Vertrag: Stand und Perspektiven, in: Neuneck, Götz/Mölling, Christian (Hg.): Die Zukunft der Rüstungskontrolle, Baden-Baden, 308–323.

Bughin, Jacques/Hazan, Eric/Ramaswamy, Sree/Chui, Michael/Allas, Tera/Dahlström, Peter/Henke, Nicolaus/Trench, Monica 2017: Artificial Intelligence. The Next Digital Frontier?, http://www.odbms.org/2017/08/artificial-intelligence-the-next-digital-frontier-mckinsey-global-institute-study/; 19.07.2019.

Cardillo, Robert 2017: GEOINT 2017 Symposium (Remarks as prepared for Robert Cardillo), https://www.nga.mil/MediaRoom/SpeechesRemarks/Pages/GEOINT-2017-Symposium.aspx; 19.07.2019.

CFTC/SEC 2010: Findings Regarding the Market Events of May 6, 2010, https://www.sec.gov/news/studies/2010/marketevents-report.pdf; 05.01.2018

Courtland, Rachel 2016: DARPA's Self-Driving Submarine Hunter Steers Like a Human, https://spectrum.ieee.org/automaton/robotics/military-robots/darpa-actuv-self-driving-submarine-hunter-steers-like-a-human; 05.01.2018.

Croft, Stuart 1996: Strategies of arms control. A History and Typology, Manchester, UK.

Daniels, Jeff 2017: Mini-nukes and Mosquito-like Robot Weapons Being Primed for Future Warfare, https://www.cnbc.com/2017/03/17/mini-nukes-and-inspect-bot-weapons-being-primed-for-future-warfare.html; 05.01.2018.

Danks, David 2014: Learning, in: Frankish, Keith/Ramsey, William (Hg.): The Cambridge Handbook of Artificial Intelligence, Cambridge, UK, 151–167.

Dillow, Clay 2016: What Happens When You Combine Artificial Intelligence and Satellite Imagery, http://fortune.com/2016/03/30/facebook-ai-satellite-imagery/; 26.11.2017.

Eilam, Eldad 2005: Reversing. Secrets of Reverse Engineering, Indianapolis, USA.

Ernest, Nicholas/Carroll, David/Schumacher, Corey/Clark, Matthew/Cohen, Kelly/Lee, Gene 2016: Genetic Fuzzy based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions, in: Journal of Defense Management 6: 1, 1–7, https://www.omicsonline.org/open-access/genetic-fuzzy-based-artificial-intelligence-for-unmanned-combat-aerialvehicle-control-in-simulated-air-combat-missions-2167-0374-1000144.pdf; 31.01.2018.

European Commission 2017: Dual-use Export Controls, http://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/index_en.htm; 19.07.2019.

Evangelista, Matthew 1988: Innovation and the Arms Race. How the United States and the Soviet Union Develop New Military Technologies, Ithaca, NY.

Farrelly, Colleen 2016: Machine Learning by Analogy, https://www.slideshare.net/ColleenFarrelly/machine-learning-by-analogy-59094152; 09.01.2018.

Franklin, Stan 2014: History, Motivations, and Core Themes, in: Frankish, Keith/Ramsey, William (Hg.): The Cambridge Handbook of Artificial Intelligence, Cambridge, UK, 15–33.

Future of Life Institute 2015: Autonomous Weapons. An Open Letter From AI & Robotics Researchers, https://futureoflife.org/open-letter-autonomous-weapons; 14.01.2018.

Goldblat, Jozef 2002: Arms Control. The New Guide to Negotiations and Agreements, London.

Goodfellow, Ian 2016: Deep Learning, Cambridge, Massachusetts, London, England.

Goodfellow, Ian J./Shlens, Jonathon/Szegedy, Christian 2015: Explaining and Harnessing Adversarial Examples, https://arxiv.org/pdf/1412.6572.pdf; 19.07.2019.

Gubrud, Mark/Altmann, Jürgen 2013: Compliance Measures for an Autonomous Weapons Convention. ICRAC Working Paper #2, https://www.icrac.net/wp-content/uploads/2018/04/Gubrud-Altmann_Compliance-Measures-AWC_ICRAC-WP2.pdf; 19.07.2019.

Gunning, David 2016: Explainable Artificial Intelligence (XAI). Broad Agency Announcement, https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf; 07.12.2017.

Hagen, Christian/Sorenson, Jeff/Hurt, Steven/Wall, Dan 2012: Software: The Brains Behind U.S. Defense Systems, https://www.atkearney.com/documents/10192/247932/SoftwareThe_Brains_Behind_US_Defense_Systems.pdf/69129873-eecc-4ddc-b798-c198a8ff1026; 19.07.2019.

Harris Cooperation 2017: ENVI Feature Extraction Module, http://www.harrisgeospatial.com/Portals/0/pdfs/HG_ENVI_FX_module_data-sheet_WEB.pdf; 26.11.2017.

Hawkins, Jeff/Blakeslee, Sandra 2004: On Intelligence, 1. Auflage, New York.

Hochmuth, Olaf 2003: Bochumer Verifikationsprojekt – Sensorstation 2000, https://www2.informatik.hu-berlin.de/~hochmuth/bvp/; 13.12.2017.

Holtom, Paul/Bromley, Mark 2011: Transit and Trans-shipment Controls in an Arms Trade Treaty. https://www.sipri.org/sites/default/files/files/misc/SIPRIBP1107a.pdf; 19.07.2019.

Human Rights Watch 2012: Losing Humanity. The Case against Killer Robots, https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots; 14.01.2018.

International Atomic Energy Agency 2018: Research and Development Plan. Enhancing Capabilities for Nuclear Verification, Wien, Österreich.

Jaccard, Nicolas/Thomas W. Rogers/Edward J. Morton/Lewis D. Griffin 2016: Automated Detection of Smuggled High-risk Security Threats Using Deep Learning, https://arxiv.org/pdf/1609.02805.pdf, 19.07.2019..

Johnson, Michael R./Paquette, Jean-Pierre/Elbez, Julien 2014: New and Emerging Trends In Satellite Imagery, https://www.iaea.org/safeguards/symposium/2014/home/eproceedings/sg2014-papers/000042.pdf; 26.11.2017.

Kahl, Martin/Mölling, Christian 2005: Die „Revolution in Military Affairs“ und die Bedingungen und Möglichkeiten für Rüstungskontrolle, in: Neuneck, Götz/Mölling, Christian (Hg.): Die Zukunft der Rüstungskontrolle, Baden-Baden, 341–353.

Kania, Elsa B. 2017: Quest for an AI Revolution in Warfare. The PLA's Trajectory from Informatized to „Intelligentized“ Warfare, https://thestrategybridge.org/the-bridge/2017/6/8/-chinas-quest-for-an-ai-revolution-in-warfare; 14.01.2018.

Keller, John 2015: DARPA TRACE Program Using Advanced Algorithms, Embedded Computing for Radar Target Recognition, http://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html; 01.12.2017.

Kleiner, Ariel/Mackey, Lester/Jordan, Michael I. 2009: Machine Learning for Improved Automated Seismic Event Extraction, https://www.ctbto.org/fileadmin/user_upload/ISS_2009/Poster/DM-02A%20%28US%29%20-%20Ariel_Kleiner%20etal.pdf; 19.07.2019.

Knight, Will 2017: The Dark Secret at the Heart of AI. No One Really Knows How the Most Advanced Algorithms Do What They Do. That could be a problem., https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/; 11.01.2018.

Krebs, Gunter 2017: KH-11/Kennen/Crystal, http://space.skyrocket.de/doc_sdat/kh-11.htm; 25.11.2017.

Kurakin, Alexey/Goodfellow, Ian/Bengio, Samy 2017: Adversarial Examples in the Physical World, https://arxiv.org/pdf/1607.02533.pdf; 19.07.2019.

Lant, Karla 2017: China, Russia and the US are in an Artificial Intelligence Arms Race, https://futurism.com/china-russia-and-the-us-are-in-an-artificial-intelligence-arms-race/; 16.11.2017.

Legg, Shane/Hutter, Marcus 2007: A Collection of Definitions of Intelligence, https://arxiv.org/pdf/0706.3639.pdf; 19.07.2019.

Li, Wei/Gauci, Melvin/Groß, Roderich 2016: Turing learning. A Metric-free Approach to Inferring Behavior and its Application to Swarms, in: Swarm Intelligence 10: 3, 211–243.

McCloskey, Paul 2017: What's AI, and what's not, https://gcn.com/Articles/2017/03/10/defining-AI.aspx; 15.11.2017.

Morisse, Tom 2017: The Next Challenges of AI Research, https://en.fabernovel.com/insights/tech-en/the-next-challenges-of-ai-research; 18.12.2017.

Müller, Harald 2000: Früherkennung von Rüstungsrisiken in der Ära der „militärisch-technischen Revolution". Ein Register für militärische Forschung und Entwicklung, Frankfurt.

Müller, Harald/Schörnig, Niklas 2006: Rüstungsdynamik und Rüstungskontrolle. Eine exemplarische Einführung in die Internationalen Beziehungen, Baden-Baden.

Müller, Vincent C./Bostrom, Nick 2016: Future Progress in Artificial Intelligence: A Survey of Expert Opinion, in: Müller, Vincent C. (Hg.): Fundamental Issues of Artificial Intelligence, Berlin, 553–571.

Neuneck, Götz/Alwardt, Christian 2008: The Revolution in Military Affairs, its Driving Forces, Elements and Complexity, https://ifsh.de/pdf/publikationen/IFAR_Working_Paper_13.pdf?asset_id=5489; 19.07.2019.

Neuneck, Götz/Mutz, Reinhard 2000: Vorbeugende Rüstungskontrolle. Ziele und Aufgaben unter besonderer Berücksichtigung verfahrensmäßiger und institutioneller Umsetzung im Rahmen internationaler Rüstungsregime, Baden-Baden.

Niemeyer, Irmgard/Ruthowski, Joshua 2016: Satellite Imagery Processing for the Verification of Nuclear Non-Proliferation and Arms Control, Bonn.

Núñez-Nieto, Xavier/Solla, Mercedes/Gómez-Pérez, Paula/Lorenzo, Henrique 2014: GPR Signal Characterization for Automated Landmine and UXO Detection Based on Machine Learning Techniques, in: Remote Sensing 6: 10, 9729–9748, http://www.mdpi.com/2072-4292/6/10/9729/pdf; 11.01.2018.

O'Neil, Cathy 2016: Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy, 1. Auflage, New York.

Osborn, Kris 2017: Air Force Upgrades F-22 Sensors, Weapons Software, https://defensesystems.com/articles/2017/03/14/f22.aspx; 03.12.2017.

Papernot, Nicolas/McDaniel, Patrick/Goodfellow, Ian/Jha, Somesh/Celik, Z. B./Swami, Ananthram 2017: Practical Black-Box Attacks against Machine Learning (Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE), Abu Dhabi, UAE.

Park, Dong H./Hendricks, Lisa A./Akata, Zeynep/Schiele, Bernt/Darrell, Trevor/Rohrbach, Marcus 2017: Attentive Explanations. Justifying Decisions and Pointing to the Evidence, https://arxiv.org/abs/1612.04757; 19.07.2019.

Parkin, Simon 2015: Killer Robots: The Soldiers that Never Sleep, http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep; 05.01.2018.

Patton, Tamara/Lewis, Jeffrey/Hanham, Melissa/Dill, Catherine/Vaccaro, Lily 2016: Emerging Satellites for Non-Proliferation and Disarmament Verification, https://nonproliferation.org/vcdnp/wp-content/uploads/2016/06/160614_copernicus_project_report.pdf; 19.07.2019.

Pilat, Joseph F. 2002: Verification and Transparency: Relics or Future Requirements?, in: Larsen, Jeffrey A. (Hg.): Arms Control. Cooperative Security in a Changing Environment, London, 79–96.

Procopio, Michael J./Young, Christopher J./Gauthier, John A. 2009: Applying Machine Learning Methods to Improve Efficiency and Effectiveness of the IDC Automatic Event Detection System, https://www.ctbto.org/fileadmin/user_upload/ISS_2009/Poster/DM-07A__US__-_Michael_Procopio_etal.pdf; 19.07.2019.

Ribeiro, Marco T./Singh, Sameer/Guestrin, Carlos 2016: „Why Should I Trust You?“. Explaining the Predictions of Any Classifier, https://arxiv.org/pdf/1602.04938.pdf; 19.07.2019.

Russell, Stuart J./Norvig, Peter 2010: Artificial Intelligence. A Modern Approach, 3. Auflage, Upper Saddle River, NJ.

Russell, Stuart J./Vaidya, Sheila/Le Bras, Ronan 2010: Machine learning for Comprehensive Nuclear-Test-Ban Treaty monitoring, in: CTBTO Spectrum: 14, 32–35; 20.11.2017.

Russia Today 2017: Kalashnikov Develops Fully Automated Neural Network-based Combat Module, https://www.rt.com/news/395375-kalashnikov-automated-neural-network-gun/; 05.01.2018.

Sauer, Frank 2016: Stopping ‘Killer Robots’: Why Now Is the Time to Ban Autonomous Weapons Systems, https://www.armscontrol.org/ACT/2016_10/Features/Stopping-Killer-Robots-Why-Now-Is-the-Time-to-Ban-Autonomous-Weapons-Systems#note04; 13.01.2018.

Scharre, Paul 2016: Autonomous Weapons and Operational Risk, https://www.files.ethz.ch/isn/196288/CNAS_Autonomous-weapons-operational-risk.pdf; 19.07.2019.

Schelling, Thomas C./Halperin, Morton H. 1961: Strategy and Arms Control, New York.

Schlosser, Eric 2013: Command and Control. Nuclear Weapons, the Damascus Accident, and the Illusion of Safety, New York, NY.

Schmidt, Hans-Joachim 2017: Hoffnungsvoller Neustart der konventionellen Rüstungskontrolle?, https://blog.prif.org/2017/07/10/hoffnungsvoller-neustart-der-konventionellen-ruestungskontrolle/; 14.01.2018.

Schörnig, Niklas 2008: Casualty Aversion in Democratic Security Provision. Procurement and the Defense Industrial Base., in: Evangelista, Matthew (Hg.): Democracy and Security. Preferences, Norms and Policy-making, London, 14–35.

Schörnig, Niklas 2015: From Quantitative to Qualitative Arms Control: The Challenges of Modern Weapons Development, in: Development and Peace Foundation/Käte Hamburger Kolleg/Centre for Global Cooperation Research (Hg.): Global Trends 2015. Prospects for World Society, 87–100.

Searle, John R. 1980: Minds, Brains, and Programs, in: Behavioral and Brain Sciences 3: 3, 417–457.

Seiffert, Udo/Abeynayake, Canicious/Jain, Lakhmi C./Tran, Minh D.-J. 2013: Detection of Targets in Characteristic GPR Sensor Data Using Machine Learning Techniques, https://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_01_2013.pdf; 19.07.2019.

Shalal, Andrea 2014: DigitalGlobe Gains U.S. Govt License to Sell Sharper Satellite Imagery, https://www.reuters.com/article/digitalglobe-imagery/digitalglobe-gains-u-s-govt-license-to-sell-sharper-satellite-imagery-idUSL2N0OR2UX20140611; 25.11.2017.

Shalev-Shwartz, Shai/Ben-David, Shai 2014: Understanding Machine Learning: From Theory to Algorithms, New York.

Sharif, Mahmood/Bhagavatula, Sruti/Bauer, Lujo/Reiter, Michael K. 2016: Accessorize to a Crime. Real and Stealthy Attacks on State-of-the-Art Face Recognition, Wien.

Shaw, Martin 2005: The New Western Way of War. Risk-transfer War and its Crisis in Iraq, Cambridge, UK.

Stevenson, Beth 2016: Analysis: Taranis Developers Reveal Test Flight Specifics, https://www.flightglobal.com/news/articles/analysis-taranis-developers-reveal-test-flight-spec-425347/; 19.07.2019.

Stocki, Trevor J./Li, Guichong/Japkowicz, Nathalie/Ungar, R. K. 2010: Machine Learning for Radioxenon Event Classification for the Comprehensive Nuclear-Test-Ban Treaty, in: Journal of environmental radioactivity 101: 1, 68–74.

Sun, Yijun/Li, Jian 2005: Adaptive Learning Approach to Landmine Detection, in: IEEE Transactions on Aerospace and Electronic Systems 41: 3, 1–9.

Sundaresan, Lalitha/Chandrashekar, S./Jasani, Bhupendra 2017: Discriminating Uranium and Copper Mills Using Satellite Imagery, in: Remote Sensing Applications: Society and Environment: 5, 27–35.

Szegedy, Christian/Zaremba, Wojciech/Sutskever, Ilya/Bruna, Joan/Erhan, Dumitru/Goodfellow, Ian/Fergus, Rob 2014: Intriguing Properties of Neural Networks, https://arxiv.org/pdf/1312.6199.pdf; 19.07.2019.

Truong, Q. B./Borstad, Gary/Saper, Ron 2005: Integration of Satellite Imagery and Other Tools in Safeguards Information Analysis, https://www.remote-sensing.aslenv.com/documents/ESARDA_paper_2005_TRUONG_etal.pdf; 19.07.2019.

Tuma, Matthias/Igel, Christian 2009: Kernel-Based Machine Learning Techniques for Hydroacoustic Signal Classification, CTBTO International Scientific Studies Conference, Wien.

UNIDIR/VERTIC 2003: Coming to Terms with Security. A Handbook on Verification and Compliance. http://www.unidir.org/files/publications/pdfs/coming-to-terms-with-security-a-handbook-on-verification-and-compliance-en-554.pdf; 19.07.2019.

US Navy 2017: MK 15 – Phalanx Close-in Weapons System (CIWS), http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2; 05.01.2018.

Yu, Xie/Jing, Meng 2017: China Aims to Outspend the World in Artificial Intelligence, and Xi Jinping Just Green Lit the Plan, http://www.scmp.com/business/china-business/article/2115935/chinas-xi-jinping-highlights-ai-big-data-and-shared-economy; 16.11.2017.

Zeiler, Matthew D./Fergus, Rob 2013: Visualizing and Understanding Convolutional Networks, https://arxiv.org/pdf/1311.2901.pdf; 19.07.2019.

ZULETZT ERSCHIENEN

PRIF REPORT

Die PRIF Reports analysieren Hintergründe politischer Ereignisse und Entwicklungen und präsentieren wissenschaftliche Forschungserbnisse in Deutsch oder Englisch.

Rähme, Sophie (2019): Kolumbiens Gewalt-Frieden. Zum Anstieg der Gewalt gegen soziale Aktivisten und Aktivistinnen seit dem Friedensabkommen 2016, PRIF Report 3/2019, Frankfurt/M.

Gromes, Thorsten (2019): A Humanitarian Milestone? NATO's 1999 intervention in Kosovo and trends in military responses to mass violence, PRIF Report 2/2019, Frankfurt/M.

www.hsfk.de/PRIF-Reports
www.hsfk.de/HSFK-Reports

PRIF SPOTLIGHT

Die PRIF Spotlights diskutieren aktuelle politische und gesellschaftliche Themen.

HSFK (2019): Kein Frieden ohne Menschenrechte. Die Verleihung des Hessischen Friedenspreises 2018 an Şebnem Korur Fincancı, PRIF Spotlight 9/2019, Frankfurt/M.

Coni-Zimmer, Melanie/Peez, Anton (2019): Deutschland im UN-Sicherheitsrat. Arria-Formel-Sitzungen als Instrument der Krisenbewältigung und -prävention, PRIF Spotlight 8/2019, Frankfurt/M.

www.hsfk.de/PRIF-Spotlights

PRIF BLOG

Auf dem PRIF Blog erscheinen Beiträge zu aktuellen politischen Fragen und Debatten der Friedens- und Konfliktforschung. Die Blogbeiträge erscheinen in loser Folge in Deutsch oder Englisch.

PRIF BLOG

https://blog.prif.org/

PRIF Reports und PRIF Spotlights sind Open-Access-Publikationen und können kostenlos auf www.hsfk.de heruntergeladen werden. Sie möchten die digitalen Ausgaben abonnieren? Bitte wenden Sie sich an: publikationen@hsfk.de.

www.facebook.com/HSFK.PRIF
www.twitter.com/HSFK_PRIF
https://blog.prif.org/

NICO LÜCK //

LERNENDE KÜNSTLICHE INTELLIGENZ IN DER RÜSTUNGSKONTROLLE

Künstliche Intelligenz, insbesondere selbstlernende KI, ist in aller Munde: auch in der Rüstung spielen solche Systeme eine zunehmend größere Rolle: Manche Waffensysteme sind bereits in der Lage, Ziele eigenständig zu identifizieren und zu bekämpfen. Dies stellt die klassische Rüstungskontrolle zunächst vor Probleme, denn ihre Verfahren sind auf die Regulierung physischer Gegenstände wie Raketen und Sprengköpfe und deren innerer Funktionsweise ausgelegt. Auch weitere wichtige Effekte einer verlässlichen Kontrolle wie Vertrauensbildung und Stabilisierung von diplomatischen Beziehungen werden erschwert. Solchen Risiken muss sich die Rüstungskontrolle stellen.

Nico Lück ist Politikwissenschaftler und hat den M.A. Internationale Studien / Friedens- und Konfliktforschung erworben. Von 2015 bis 2018 war er am Leibniz-Institut Hessische Stiftung Friedens- und Konfliktforschung (HSFK) im „EU Non-Proliferation and Disarmament Consortium“ beschäftigt. Aktuell arbeitet er bei der Deutschen Gesellschaft für Internationale Zusammenarbeit (GIZ).

ISBN 978-3-946459-46-0

PRIF Peace Research Institute Frankfurt
HSFK Leibniz-Institut Hessische Stiftung Friedens- und Konfliktforschung
Leibniz Gemeinschaft